EXPLORING DISTANCE-AWARE UNCERTAINTY QUANTIFICATION FOR REMOTE SENSING IMAGE CLASSIFICATION

Christoph Koller [1,2], *Peter Jung* [2,3], *and Xiao Xiang Zhu* [1]

[1] Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), Germany
[2] German Aerospace Center (DLR), Germany
[3] Communications and Information Theory Group, Technical University of Berlin, Germany

ABSTRACT

Deep Learning models for classification often suffer from overconfidence, which naturally results in poor predictive uncertainty estimates. To overcome this, many calibration techniques have been established. These techniques operate on the labels or the output space of the network but ignore the input image space. A recently proposed approach considers the distances between different network inputs explicitly and theoretically propagates the distances through the network. The resulting predictive uncertainties of the model are then able to better reflect these distances. We test this approach in the context of remote sensing image classification for land use. To evaluate the predictive uncertainties, we set up an Out-of Distribution (OoD) detection framework based on class separation.

Index Terms— Land Use, Classification, Uncertainty Quantification, Out-of-Distribution (OoD), OoD Detection, Residual Network, Spectral Normalization

1. INTRODUCTION

When performing a classification task, we rely on a model output for the classification decision. Often the model yields a predictive probability which indicates its confidence. Preferably, we would also like to receive a notion of the model's uncertainty hinting at data points the model struggles to classify. This uncertainty, if trustworthy, is beneficial for downstream applications such as out-of-distribution (OoD) detection and active learning. Generally, we can distinguish between deterministic and probabilistic approaches, where the latter includes approaches with both multiple models and forward passes [1]. While probabilistic approaches generally achieve good performance results, deterministic approaches are less complex and there promising for onboard architectures. They have shown competitive results in some machine learning benchmarks [2] and good performance for the task of OoD detection in Remote Sensing (RS) [3].

Whereby many approaches rely on heuristic arguments, we here follow a deterministic approach [4] with theoretical justification. The approach combines residual networks with a Gaussian process output layer on top. As another peculiarity, the weight matrices are spectrally normalized. The core idea is shown in Figure 1. By using this construction, the authors give theoretical justification for a sensitivity towards distances in the model's input space. In the remainder of the paper, the approach is briefly summarized and experiments based on OoD detection within remote sensing image classification are conducted.

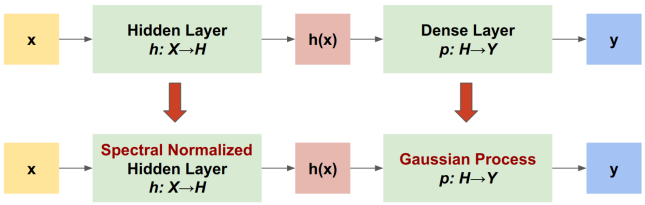


Fig. 1: General framework for distance-aware uncertainty quantification [5] with residual network architectures. Notation and discussion given in Section 2.

2. METHODOLOGY

Following [4], in order to have representative uncertainties, a sense of awareness of the input data needs to be established. Once given this awareness, measures of uncertainty based on the predictive distribution adequately reflect the distance between instances in the input space. The authors describe the *input distance awareness* by requiring an uncertainty measure $u : \mathcal{F}(\mathcal{X}) \to \mathbb{R}_+$, operating on the output space of the network $\mathcal{F}$, where $\mathcal{X}$ denotes the input space of the network. The uncertainty measure needs to reflect the distance between a new point $x \in \mathcal{X}$ and the training data $\mathcal{X}_{\text{train}} \subset \mathcal{X}$ with respect to a suitable metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, e.g., of the form $d_X(x_1, x_2) = \|x_1 - x_2\|_{\mathcal{X}}$, $x_1, x_2 \in \mathcal{X}$. More precisely, it is assumed that u can be represented with a monotonically increasing function $v : \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

$$\forall x \in \mathcal{X} : \; u(x) = v(\mathbb{E}_{x' \sim \mathcal{X}_{\text{train}}} d(x, x')) \tag{1}$$

The authors in [4] state that in a classification setting, for a logit of a neural network $\text{logit}(x) = g \circ h(x)$ to be *input distance aware*, the output mapping g is supposed to be *distance aware* and the hidden mapping h (almost) *distance preserving*. Distance preservation is essentially equivalent to a bi-Lipschitz constraint on the hidden mapping h, i.e., there exists constants $0 < L_1 \leq 1 \leq L_2$ such that

$$\underbrace{L_1 \cdot d_X(x, x') \leq \overbrace{d_X(h(x), h(x'))}^{\text{Feature space smoothness}} \leq L_2 \cdot d_X(x, x')}_{\text{Input distance sensitivity}} \quad (2)$$

holds for all $x, x' \in \mathcal{X}$. When considering residual network architectures (ResNets) [6], the hidden mapping is of the form $h(x) = h_l \circ h_{l-1} \circ ... \circ h_1(x)$, where $h_j(x) = x + \gamma_j(x)$, $j = 1, \ldots, l$ are the *residual blocks*. The bi-Lipschitz condition is then fulfilled if and only if the residual mappings $\gamma_j(x)$ of the residual blocks are α-Lipschitz for $0 < \alpha < 1$ on $\mathcal{X}$. This condition is already sufficient since the concatenation of α-Lipschitz functions is again α-Lipschitz [7]. Given the l residual blocks, the Lipschitz constants of the concatenation result in

$$0 < L_1 = (1 - \alpha)^l < 1 < L_2 = (1 + \alpha)^l < \infty\,. \quad (3)$$

If the dataset would span the entire $\mathbb{R}^n$ equipped with the Euclidean norm (and Bi-Lipschitz continuity is supposed to hold true on this entire space), the condition $\alpha < 1$ corresponds to controlling the spectral norm (largest singular value λ) of the weight matrix W_j of γ_j. In other non-trivial cases however, one has to resort to approximations anyway. For example, a simple and effective technique to approximate λ during training can be done by the power iteration [8]. In particular, for each layer $j = 1, \ldots, l$, we update the weight matrix via $W_j \leftarrow c \cdot W_j / \hat{\lambda}$ if $c < \hat{\lambda}$. Here, $\hat{\lambda}$ is the approximation for λ obtained by the power iteration and c is a further hyper-parameter to practically control the exact upper bound depending on the input data.

For the output mapping g to be distance-aware, the authors in [4] propose to model g as a Gaussian process (GP) [9] on the hidden mapping output space $\mathbb{H} := \{h(x) : x \in \mathcal{X}\}$ specified by a mean function $m(h)$ and a covariance function $k(h, h')$. The latent process is assumed to be uninformative a priori with a mean of 0 and the input awareness is accounted for by placing a radial basis function (RBF) kernel. All hidden representations of the input data are then exposed to the process, after this the posterior gets optimized. Due to computational and analytical intractability, the authors propose (i) a Random Fourier Feature expansion [10] of the initial GP followed by (ii) a Laplace approximation for the posterior. We refer the interested reader to [4].

Generally, the posterior GP easily calculates an uncertainty via the posterior variance. Since the posterior here is only approximated, the predictive uncertainty score can be e.g. calculated via the *Dempster-Shafer metric (DSM)* (proposed by the authors in [4]):

$$u(x) = \frac{K}{K + \sum_{k=1}^{K} \exp(g_k(x))} \quad (4)$$

where $g_k(x)$ denotes the approximation of the k^{th} logit via the posterior GP and K equals the number of classes. Alternatively, we can treat the value of one minus the maximum softmax probability (MSP) of the prediction as an uncertainty measure.

3. EXPERIMENTAL RESULTS & VALIDATION

Within the task of remote sensing image classification, we focus on land use classification making use of the benchmark dataset *Eurosat* [11]. The dataset contains 27,000 labeled Sentinel-2 satellite images, equally distributed among the 10 land use classes. The classes are well distinguishable and deep learning models have been shown to perform well on the classification task. In order to set up an evaluation framework for the proposed uncertainty quantification approach SNGP, the dataset is divided into building classes (classes 4 and 7, or industrial buildings and residential buildings, respectively) and vegetation classes (remaining classes). As can be seen in Figure 2, the classes are generally well separable into ID and OoD, but overall semantic similarities remain. While the network gets trained according to the SNGP approach on one of the subsets, this subset is considered in-distribution (ID), while the other subset is used for uncertainty evaluation and is considered OoD.

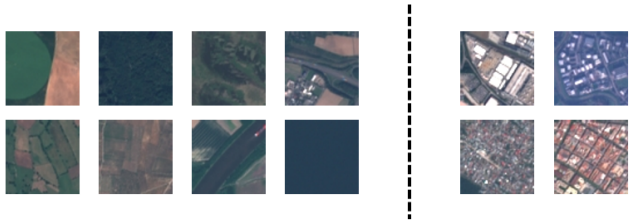


Fig. 2: Example images from the Eurosat [11] dataset, split into ID and OoD.

The experiments in this work are largely based on the code implementation from [12]. For the core network, we consider a ResNet50 [6] and a WideResnet [13] with a depth of 28 and a single filter, as the use of multiple filters and the same depth was found to be inferior. The networks were trained with a piecewise constant learning rate initiated at 0.1 (0.01 for ResNet50) and decaying every 5 epochs with a ratio of 0.1. Training was performed via Stochastic Gradient Descent (SGD) with Nesterov momentum of 0.9 for 20 epochs. Regarding the spectral normalisation, a single power iteration was performed and the aforementioned hyper-parameter was set to $c = 6$ as proposed by the authors in [4]. The Gaussian process was initiated with a hidden dimension of 1024, a

ID Data	OoD Data	Method	NLL ↓	AUROC ↑		AUPR ↑	
				DSM	MSP	DSM	MSP
Vegetation classes	Building Classes	WideResNet	0.93	0.32	0.65	0.32	0.59
		WideResNet-SN	0.94	0.17	0.74	0.34	0.66
		WideResNet-GP	1.10	0.26	0.65	0.36	0.60
		WideResNet-SNGP	1.38	0.21	0.63	0.35	0.57
Vegetation classes	Building Classes	ResNet50	3.27	0.50	0.62	0.55	0.55
		ResNet50-SN	4.24	0.43	0.42	0.42	0.41
		ResNet50-GP	3.82	0.65	0.45	0.54	0.44
		ResNet50-SNGP	3.58	0.52	0.43	0.59	0.43

Table 1: Experimental results on the Eurosat [11] dataset. NLL = negative log-likelihood on the ID test set, AUROC = area under receiver operator curve, AUPR = area under precision recall curve, DSM = Dempster-Shafer metric, MSP = 1 - maximum softmax probability. AUROC and AUPR are based on a binary classifier distinguishing ID and OoD data by means of the respective predictive uncertainty quantity (OoD samples are expected to have higher uncertainties than ID samples).

RBF kernel with a bias of 0 and a length-scale parameter of 1 was used. For the kernel amplitude a value of 15 was chosen empirically in the case of the WideResNet, and a value of 1 in the case of the ResNet50. All other hyper-parameters were left unchanged (based on experiments with Cifar-10 from the original authors in [12]).

Table 1 shows quantitative results of the earlier described experiments. For each network, the plain network is compared with the addition of spectral normalization (-SN), with the use of a Gaussian process classification head (-GP), and with the full approach explained in [4] (-SNGP). For the OoD detection with WideResNets, a few interesting aspects can be highlighted. Overall, the Dempster-Shafer metric does not seem to adequately capture the uncertainty within the logits, since there lies a large gap between the AUROC and AUPR of the DSM and the MSP for all configurations. Also, surprisingly the test loss does not always correlate with the OoD detection performance, as the plain WideResNet has the overall lowest loss but lacks behind the WideResNet-SN for the AUROC and AUPR (both for DSM and MSP). Including the spectral normalisation of the weight matrices seems to add value to almost every configuration, with the AUROC via DSM being the only exception for the WideResNet. The Gaussian process, on the other hand, has mixed effects. Note that the GP configuration comes with many hyper-parameters, which are to be controlled and which can greatly impact the model performance.

In the case of the ResNet50, the first point worth mentioning is the vastly shifted scale on which the NLL of the test set lies. This cannot easily be explained, but possible reasons can be given by the higher complexity of the model interchanging with the complexity of spectral normalisation and the Gaussian process, as well as the limited amount of resources spent on hyper-parameter optimization within this work. Surprising at the same time is the strong performance of the Dempster-Shafer metric. While the spectral normalisation here seems to have mostly negative impact on the performance, the use of the Gaussian process output layer improves the AUROC in combination with the DSM. The full SNGP model can only lead to a performance gain in a single configuration, namely for the AUPR via DSM.

Figure 3 displays the binned predictive uncertainties and ROC curves for the two networks considered. Note that while the resulting AUROC is almost identical, the predictive uncertainties for both ID and OoD samples largely differ. The WideResNet yields a wide range of predictive uncertainties, therefore especially for higher predictive uncertainties, it is hard to differentiate ID and OoD samples. The ResNet50, on the other hand, predicts generally very low uncertainty samples for ID data, therefore it becomes easier to identify OoD samples the higher the predictive uncertainty is.

4. CONCLUDING REMARKS

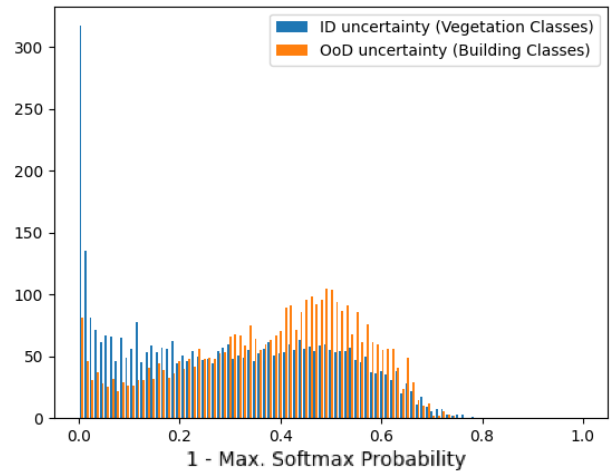
The methodology stated in this work cleverly combines spectrally normalized weight matrices with a distance-aware Gaussian process as output layer. By doing so, theoretical guarantees can be given on the input distance preservation throughout the network, which in turn yield reliable predictive uncertainty estimates. Although the structural changes of the SNGP approach to an existing network are relatively few, the complexity of the optimization is manifold. We believe there exists unleashed potential for the application in remote sensing image classification which we leave for future works to explore.

5. REFERENCES

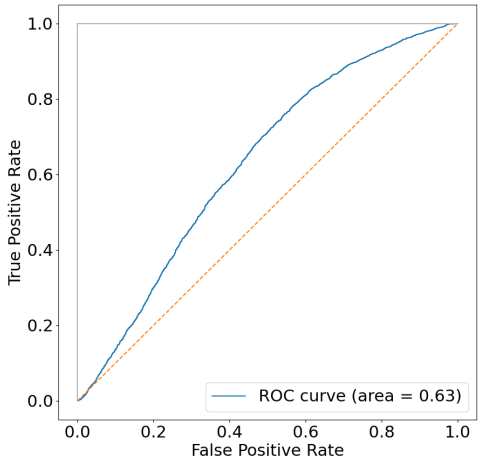
[1] J. Gawlikowski, C. R. N. Tassi, et al., "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[2] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *ICML*, 2020.

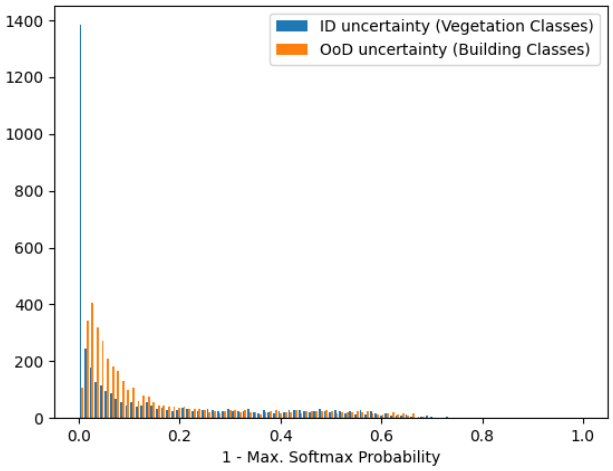
[3] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, 60, 2022.



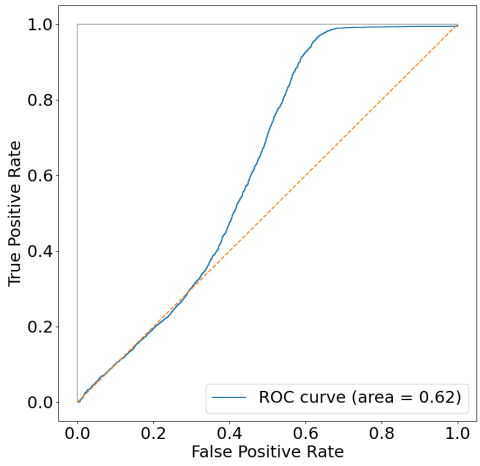
(a) WideResNet



(b) WideResNet



(c) ResNet50



(d) ResNet50

Fig. 3: Exemplary (a) & (c) Histograms of predictive uncertainty scores via MSP and (b) & (d) ROC curves for binary OoD classifier. The two plots are showing the interdependence between the grouped uncertainty quantities and the performance of the binary OoD detector highlighting two vastly different challenges with the approach.

[4] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *NeurIPS*, 2020.

[5] T. T. Authors., "Uncertainty-aware deep learning with sngp tensorflow tutorial," `https://github.com/tensorflow/docs/blob/master/site/en/tutorials/understanding/sngp.ipynb`, 2021, [Online, Accessed: 2023-01-10].

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[7] P. L. Bartlett, S. N. Evans, and P. M. Long, "Representing smooth functions as compositions of near-identity functions with implications for deep network optimization," *arXiv preprint arXiv:1804.05012*, 2018.

[8] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," in *ICML*, 2019.

[9] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006.

[10] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *NeurIPS*, 2007.

[11] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, 12, no. 7, 2019.

[12] Z. Nado, N. Band, M. Collier, et al., "Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning," *arXiv preprint arXiv:2106.04015*, 2021.

[13] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.