

# OUT-OF-DISTRIBUTION DETECTION IN SATELLITE IMAGE CLASSIFICATION

**Jakob Gawlikowski**

Institute of Data Science  
German Aerospace Center, Jena, Germany  
jakob.gawlikowski@dlr.de

**Sudipan Saha**

Data Science in Earth Observation  
Technical University of Munich, Germany  
sudipan.saha@tum.de

**Anna Kruspe**

Data Science in Earth Observation  
Technical University of Munich, Germany  
anna.kruspe@tum.de

**Xiao Xiang Zhu**

Remote Sensing Technology Institute  
German Aerospace Center, Weßling, Germany  
xiaoxiang.zhu@dlr.de

## ABSTRACT

In satellite image analysis, distributional mismatch between the training and test data may arise due to several reasons, including unseen classes in the test data and differences in the geographic area. Deep learning based models may behave in unexpected manner when subjected to test data that has such distributional shifts from the training data, also called out-of-distribution (OOD) examples. Predictive uncertainty analysis is an emerging research topic which has not been explored much in context of satellite image analysis. Towards this, we adopt a Dirichlet Prior Network based model to quantify distributional uncertainty of deep learning models for remote sensing. The approach seeks to maximize the representation gap between the in-domain and OOD examples for a better identification of unknown examples at test time. Experimental results on three exemplary test scenarios show the efficacy of the model in satellite image analysis.

## 1 INTRODUCTION

Deep learning has revolutionized the field of remote sensing in the last few years Ball et al. (2017); Mou et al. (2021); Saha et al. (2019). Most of the satellite image analysis approaches assume that test data is similarly distributed as the training data on which the model is trained. However, this assumption rarely holds in practice. Remote sensing deals with a large number of acquisition sensors operating across a variety of different geographies. Moreover, some landscape classes seen be seen in only some geographic areas. Deep learning models are likely to fail or behave in an unexpected way when faced with open-set classes. A deep model trained on images from agricultural area will likely fail when asked to predict urban images comprising unseen classes. Similarly, deep models behave in unexpected way when fed with data from seen classes but with considerable geographic variation. For example, European and Asian urban areas exhibit significantly different semantics and a model trained on one may likely fail on the another, forcing to use geography-wise different models Saha et al. (2020). When deep learning based systems fail, they do not provide sufficient cue to the user and can give a wrong prediction, yet with high confidence. To address this issue, predictive uncertainty estimation has recently emerged as a research topic in the machine learning community Malinin & Gales (2018). Uncertainty estimation informs users about the confidence on a prediction, thus gives a hint on the reliability of such systems and possible weaknesses.

Deep learning based classification models are prone to predictive uncertainties from three different sources Malinin & Gales (2018): model or epistemic uncertainty, data or aleatoric uncertainty, and distributional uncertainty. In remote sensing distributional uncertainty may arise due to various reasons, as unseen classes, geographic differences, and sensor differences. Considering its high relevance in satellite image analysis, our work focuses on distributional uncertainty Gal (2016).

Our work is based on a Dirichlet Prior Network (DPN) that separately models different aforementioned uncertainty types. The Dirichlet distribution is a distribution over the categorical distribution,

i.e. it can model uncertainty on a soft-max output of a classification model. DPNs separate in-distribution and OOD examples by producing sharp Dirichlet distributions for in-domain examples (low deviation in the softmax output) while producing flat Dirichlet distributions for OOD ones (high deviation in the softmax output) Malinin & Gales (2018). In particular, we base our work on an extension of the DPN classifier Nandy et al. (2020) that focuses on increasing the representation gap between in-domain and OOD examples. We experimentally show that the proposed approach is able to detect OOD examples in remote sensing images, thus improving the reliability and robustness of deep learning based models in remote sensing. To the best of our knowledge this is the first work that specifically addresses out-of-distribution detection in remote sensing.

## 2 DPN FOR SATELLITE IMAGE ANALYSIS

In satellite image classification, images  $x$  and their corresponding labels  $y$  can be characterized using their distribution  $p(x, y)$ . In practice, we only have a finite data set  $\mathbb{D} = \{x_j, y_j\}_{j=1}^N$  corresponding to the distribution  $p(x, y)$ . Since the training data is a random subset and the training process is also affected by randomness, Bayesian neural networks model the parameters  $\theta$  of a neural network as a random variable. For a classifier with parameters  $\theta$  the predictive uncertainty on a prediction  $\omega$  is then given by  $p(y = \omega | x^*, \mathbb{D}) = p(y = \omega | x^*, \theta) p(\theta | \mathbb{D})$ .

The sources of predictive uncertainty Malinin & Gales (2018) can be broadly categorized into the following three categories: epistemic or model uncertainty, aleatoric or data uncertainty, and distributional uncertainty. Distributional uncertainty is likely in remote sensing due to differences caused by new classes in the target data, geographic shift, and multi-sensor differences. Approaches as Bayesian Neural Networks and deep ensembles consider the distributional uncertainty as part of the epistemic uncertainty. These approaches seek to explicitly predict the aleatoric uncertainty and to quantify the epistemic uncertainty by performing several predictions with different model parameters Lakshminarayanan et al. (2017).

Dirichlet distributions are popularly used as a prior distribution in Bayesian learning. Malinin and Gales Malinin & Gales (2018) proposed Dirichlet Prior Networks (DPN) that efficiently mimic the behavior of Bayesian networks by parameterizing a Dirichlet distribution over the categorical distribution given by a soft-max classification output. Convenient to remote sensing applications, any neural network with soft-max activation can be considered as a DPN. A Dirichlet distribution over  $K$  classes is characterized by concentration parameters  $\{\alpha_1, \dots, \alpha_K\} > 0$ . For a DPN the concentration is given by the exponentials of the network’s logit values  $z$ ,

$$\alpha_k = \exp(z_k(x^*)). \tag{1}$$

The sum of the concentrations  $\alpha_0 = \alpha_1 + \dots + \alpha_K$  is called the precision of the distribution. The larger the precision, the sharper is the Dirichlet distribution.

For in-domain samples where the classifier is confident, DPNs aim to produce uni-modal distribution at the corner of the solution simplex with the correct class (Figure 1(a)) Malinin & Gales (2018). For in-domain samples with high data uncertainty DPNs aim to produce a sharp distribution at the center (Figure 1(b)) and for OOD data a flat distribution (Figure 1(c)). However, for in-domain examples with high aleatoric uncertainty among multiple classes, DPNs could also produce flat Dirichlet distributions Nandy et al. (2020), what often leads to representations which are indistinguishable from OOD examples. To overcome this, Nandy et al. Nandy et al. (2020) proposed the  $DPN^-$  approach.  $DPN^-$  aims at learning a sharp multi-modal distribution ( $\alpha_0 \ll 1$ ) instead of a flat uni-modal distribution for OOD examples. Additionally, Nandy et al. chose DPN parameters in a way, that the loss simplifies to the cross-entropy plus a precision regularization term.

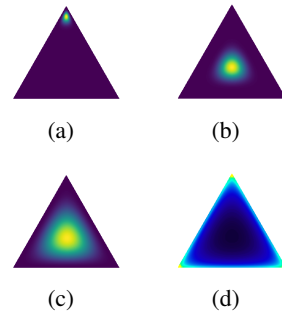


Figure 1: Different desired predictive uncertainties shown over the simplex (cf. Nandy et al. (2020)): (a) In-domain confident, (b) In-domain aleatoric uncertainty, (c) OOD (with DPN Malinin & Gales (2018)), (d) OOD (with  $DPN^-$  Nandy et al. (2020)).

The precision regularization is achieved by introducing a bounded regularization term

$$\alpha'_0 = \frac{1}{K} \sum_{k=1}^K \text{sigmoid}(z_k(x))$$

as a regularizer along with the cross-entropy loss. This gives the following two loss formulations for in-domain and OOD examples:

$$\mathcal{L}_{in}(\theta; \lambda_{in}) := \mathbb{E}_{P_{in}(x,y)} [-\log p(y|x, \theta) - \lambda_{in} \alpha'_0] \quad (2)$$

and

$$\mathcal{L}_{out}(\theta; \lambda_{out}) := \mathbb{E}_{P_{out}(x,y)} [\mathcal{H}_{ce}(\mathcal{U}; p(y|x, \theta)) - \lambda_{out} \alpha'_0] . \quad (3)$$

$\mathcal{U}$  denotes the uniform distribution over all classes,  $\mathcal{H}_{ce}$  denotes the cross-entropy function, and the precision is controlled by two hyper-parameters  $\lambda_{in} > 0$  and  $\lambda_{out} < 0$ . The combined loss-function is given by

$$\mathcal{L}(\theta; \gamma, \lambda_{in}, \lambda_{out}) = \mathcal{L}_{in}(\theta, \lambda_{in}) + \gamma \mathcal{L}_{out}(\theta, \lambda_{out}), \quad (4)$$

where in-domain and OOD samples are balanced by  $\gamma > 0$ .

For in-domain examples which are confidently predicted, the cross-entropy loss maximizes the logit value of the correct class. However, for in-domain samples with aleatoric uncertainty, the optimizer maximizes  $\text{sigmoid}(z_k(x))$  for all classes, thus yielding a flatter distribution. By choosing  $\lambda_{out} < 0$ ,  $\text{DPN}^-$  produces uniform negative values for  $z_k(x^*)$  for an OOD example  $x^*$ . This leads to  $\alpha_k \ll 1$  for all  $k = 1, \dots, K$ , and thus an OOD sample yields a sharp multi-modal Dirichlet distribution with uniform weights at each corner of the simplex (Fig 1(d)). Figures 1(b) and 1(d) are more distinct over the simplex, making the OOD samples easier distinguishable from the in-domain ones. In Figure 2 a visualization of the training process of  $\text{DPN}^-$  is given.

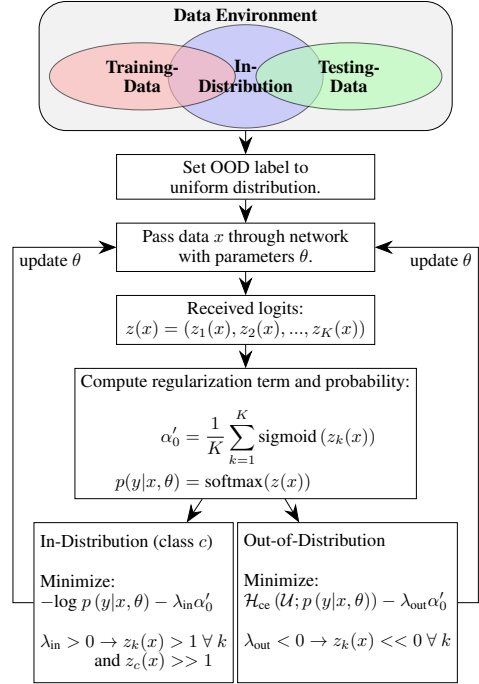


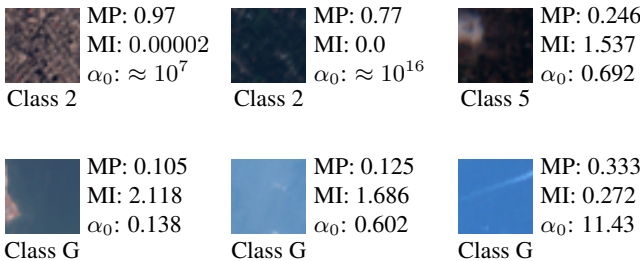
Figure 2: Visualization of the training procedure for the considered  $\text{DPN}^-$  network.

### 3 EXPERIMENT AND RESULTS

		DPN <sup>-</sup> network		Binary Classifier	
		Testing Data Set	Left out 10% of Training Set	Testing Data Set	Left out 10% of Training Set
Test Case 1	Max. Prob.	95.51±1.63	98.66±0.37	90.67±1.10	91.87±1.76
	Mutual Info	96.28±0.57	99.24±0.32	-	-
	$\alpha_0$	96.26±0.51	99.23±0.33	-	-
Test Case 2	Max. Prob.	73.99±3.59	87.88±2.54	60.31±4.53	73.79±4.58
	Mutual Info	81.81±1.68	93.85±1.06	-	-
	$\alpha_0$	85.15±1.94	95.01±0.88	-	-
Test Case 3	Max. Prob.	83.15±3.46	92.27±2.88	53.73±9.86	86.42±4.93
	Mutual Info	87.03±1.21	95.62±2.80	-	-
	$\alpha_0$	86.94±1.17	95.53±2.75	-	-

Table 1: AUROC scores of the  $\text{DPN}^-$  and a binary classifier baseline network. The scores are based on maximum probability, mutual information, and precision for the  $\text{DPN}^-$ . For the binary classifier, only the maximum probability is considered, since  $\alpha_0$  is related to the Dirichlet distribution and mutual information can not be used for a binary variable. The results are given as mean and standard deviation of five runs.

In order to evaluate the gap between in-domain and OOD samples we use the same measures as in Nandy et al. (2020), namely mutual information, maximum probability, and the precision  $\alpha_0$ . The general performance is characterized by *area under the receiver operating characteristic (AUROC)* scores based on these three measures.



*Test dataset:* We use the So2Sat LCZ42 dataset Zhu et al. (2019) for evaluating the OOD detection performance. The dataset consists of local climate zone (LCZ) labels of approximately half a million Sentinel-2 patches. Note that Sentinel-2 satellite images are significantly different from natural images (used in computer vision) having 13 spectral bands and 10 m/pixel spatial resolution. The local climate zones are described by 17 classes, 1-10 corresponding to urban areas, A-F corresponding to non-urban areas, and G corresponding to water body. We performed our experiments using following combinations:

1. Urban classes as in-domain data, non-urban ones as OOD data during training, and water body as OOD data during test.
2. Red channels (corresponding to all 17 classes) as in-domain, green channels as OOD during training, and blue channels as OOD during test.
3. Urban and vegetation classes as in-domain, rock and pavement as OOD during training, and water as OOD during test.

Figure 3: Visualization of example samples from the left out 10% of the training set of the So2Sat LCZ42 data set. The results are based on the  $DPN^-$  network trained on urban (in-distribution) and vegetation (out-of-distribution) samples. One can clearly see the differences in the metrics. The two examples on the right side do not fit well into our assumptions, possibly caused by the clear edge in the water image and the blur in the urban image.

*Deep architecture:* We used five sequential layers with 32, 64, 64, and 128 convolutional filters of size 3x3 each, followed by a dense layer of size 256. After each convolution layer, batch normalization is applied. The networks are trained for 200 epochs.

*Comparison methods:* We consider a binary classifier trained to separate in-domain and OOD data. We evaluate the performance on a left-out 10% subset of the training set (evaluation on seen regions) and on the OOD samples from unseen regions.

*Results:* In Table 3 the results based on 5 runs for each setting are presented and in Figure 3 six examples are shown. The  $DPN^-$  network clearly outperforms the binary classifier in separating in-domain and OOD examples on seen and unseen regions. The use of mutual information or the precision value contributes to increase the AUROC scores for the  $DPN^-$  network for all test instances. Among the different considered cases, separating urban and vegetation classes is clearly most trivial, while the exclusion of single classes, as in test case 3, is significantly difficult. However,  $DPN^-$  still perform satisfactorily for this task.

## 4 CONCLUSION

In this paper, we quantified distributional uncertainty in deep learning models for satellite image analysis. We tested the method on the So2Sat LCZ42 dataset considering open set classes and selected bands as OOD. Satellite images are significantly different from the natural images dealt in computer vision. It is important to understand predictive uncertainty in context of satellite image analysis and our work is a first step towards it.

## REFERENCES

John E Ball, Derek T Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote*

- Sensing*, 11(4):042609, 2017.
- Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3):4, 2016.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.
- Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.
- Lichao Mou, Sudipan Saha, Yuansheng Hua, Francesca Bovolo, Lorenzo Bruzzone, and Xiao Xiang Zhu. Deep reinforcement learning for band selection in hyperspectral image classification. *arXiv preprint arXiv:2103.08741*, 2021.
- Jay Nandy, Wynne Hsu, and Mong Li Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *arXiv preprint arXiv:2010.10474*, 2020.
- Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019.
- Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Change detection in image time-series using unsupervised lstm. *IEEE Geoscience and Remote Sensing Letters*, 2020.
- Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, et al. So2sat LCZ42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019.