COMPACT FEATURE REPRESENTATION FOR UNSUPERVISED OOD DETECTION

Sudipan Saha[1], *Jakob Gawlikowski*[1,2], *Jay Nandy*[3], *Xiao Xiang Zhu*[1,4]

Data Science in Earth Observation, Technical University of Munich, Ottobrunn, Germany[1]
German Aerospace Center (DLR), Jena, Germany[2]
National University of Singapore, Singapore[3]
Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany[4]

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884481

ABSTRACT

Distributional mismatch between training and test data may cause the remote sensing models to behave in unpredictable manner, thus reducing the trustworthiness of such models. Most existing methods for out-of-distribution (OOD) detection rely on availability of OOD samples during training. However, access to OOD data during training is counter intuitive and may be impractical sometimes. Considering this, we propose an unsupervised OOD detection model that does not require training OOD data. The proposed method works by projecting the in-domain samples as a union of 1-dimensional subspaces. Due to the compact feature representation of in-domain samples, OOD samples are less likely to occupy the same feature space, thus they are easily identified. Experimental results demonstrate the capability of the proposed method to detect OOD samples.

Index Terms— Uncertainty, out-of-distribution, trustworthiness, deep learning, unsupervised learning, remote sensing

1. INTRODUCTION

There has never been a time before with such abundant remote sensing data, offering great potential for the data science enabled methods to advance understanding of many environmental problems. Most such methods are based on deep learning and have shown excellent performance in almost all remote sensing tasks, including image classification, change detection, and fusion [1, 2]. However, deep learning methods are data-hungry and require abundant amount of training data characterizing the target distribution. Deviation of test data from training data leads to significant performance decline. Such deviation may be caused by geographical shift, sensory shift, and presence of unseen classes. In addition to mere performance decline, shift in target data distribution may also lead the model to produce wrong prediction, however without giving any cue to the user about possible incorrect prediction. This may cause debacle in time-bound applications, e.g., leading rescue teams to wrong locations when identifying destructed buildings immediately after a disaster.

Predictive uncertainty estimation has recently emerged as research topic in the machine learning community [3, 4, 5]. Few works have adopted uncertainty estimation in context of remote sensing [6]. However, existing distributional uncertainty estimation works in remote sensing assume that the deep learning model has access to out-of-distribution (OOD) data while training the model and they introduce additional terms in the loss function while training [6]. We observe two pitfalls of such approach. Firstly, the access to OOD data during training is counter-intuitive to the problem of OOD detection and access to such OOD training data may be limited by different practical constraints. Secondly, introducing additional loss term significantly deviates the training objective from its original task, i.e., to obtain satisfactory classification. To alleviate these constraints, few works in the literature have explored unsupervised detection of OOD samples, i.e., without using any OOD sample during training phase [7, 8]. One straightforward approach is to use the softmax value as an indicator of OOD detection [9, 10]. However, mere softmax scores are not reliable OOD predictor [7]. Further advancing this concept, [11] postulates that unsupervised OOD detection performance can be improved by constraining the representation of in-distribution samples in the feature space. Following this, we propose an OOD detection model for remote sensing classification tasks that work by projecting the in-distribution samples into union of 1-dimensional subspaces. Due to the compact representation of the in-distribution samples in the feature space, OOD samples are much less likely to occupy the same region as them.

The contributions of this work are as follows:

1. We introduce the concept of projecting in-distribution samples into 1-dimensional subspaces in remote sensing. Our work is one of the first works in remote sensing on unsupervised OOD detection.

2. We experimentally show the effectiveness of the proposed method on open set recognition task in UC Merced dataset [12].

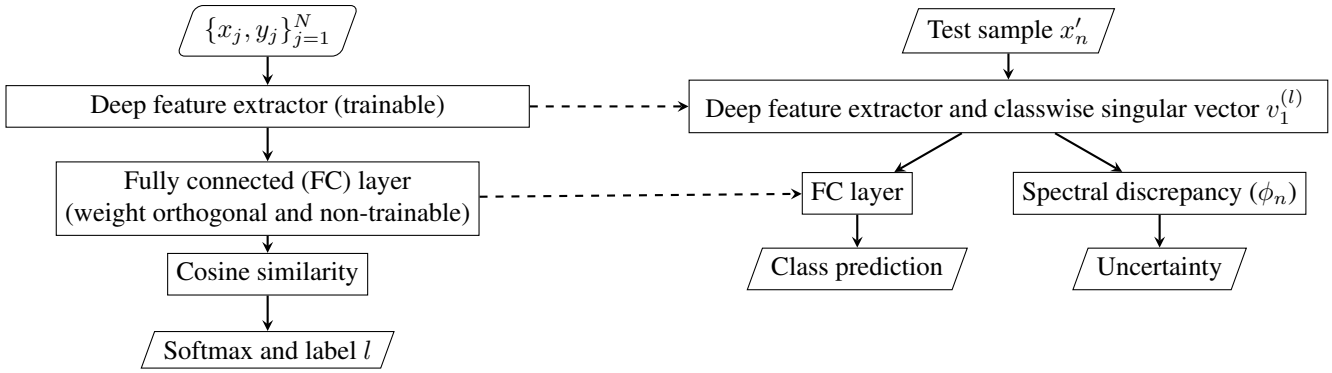

Fig. 1. Proposed unsupervised OOD detection framework. The left hand side denotes the training process with only in-domain samples while the right hand side shows the OOD detection process on test samples.

2. PROPOSED METHOD

A dataset of labeled remote sensing images $\mathcal{D} = \{x_j, y_j\}_{j=1}^N$ (labels belonging to L classes) can be characterized by their underlying probability distribution $p(x, y)$. Uncertainties caused by finite size of the dataset are defined as model/epistemic uncertainty [3]. Data/aleatoric uncertainty arises from class overlap, label noise, and other complexities in the data distribution [5]. On the other hand, distributional uncertainty arises from the differences in training data distribution ($p(x, y)$) and test data distribution ($p'(x, y)$) [6]. Distributional uncertainty is very likely in remote sensing due to presence of unseen classes in test data, i.e., not belonging to L classes seen during training. Other sources of distributional shift include geographic and sensory differences. While previous remote sensing OOD detection works use OOD samples (i.e., samples not belonging to $p(x, y)$) during training [6], in this work we propose to exclude use of any such OOD samples during training. Our goal is to train a neural network for classification using D, such that at the test time the classifier produces classification label and also an additional output on whether the test sample is OOD. Figure 1 shows a flowchart of the proposed method.

2.1. Intraclass compactness during training

OOD detection can be improved by representing the known classes present in in-domain training data in a compact feature space, e.g., union of 1-dimensional subspaces [11]. The idea is that if the known classes seen during training lie on a compact feature space, the probability that the OOD samples also lie on the same feature spaces reduces drastically. A deep network can be thought to be composition of two different entities: a feature extractor network (all layers before the last fully connected layer) that generates a feature vector f_n from an input image x_n and the last fully connected layer that maps the feature vector f_n to the L classes using weights w_l ($l = 1, .., L$). During training, cosine similarity can be employed to make the feature vectors of each known classes to lie on 1-dimensional subspace. Let us assume, the cosine similarity between the feature vector f_n and the weights w_l is given by $cos(\theta_{ln})$. Then the probability of membership of the feature vector n in class l is given by:

$$p_{ln} = \frac{exp(|cos(\theta_{ln})|)}{\sum_j exp(|cos(\theta_{jn})|)} \tag{1}$$

Using p_{ln} along with cross-entropy loss during training ensures that the feature vectors of each class l are aligned to their corresponding weight vector w_l [11].

2.2. Maximizing interclass separation during training

While the use of cosine similarity ensures intraclass compactness, we also need to maximize the interclass separation of the known classes present in in-domain training data. In other words, interclass similarity needs to be decreased among different classes in terms of cosine similarity. This can be ensured by enforcing w_l to be orthogonal to each other [11]. This can be achieved by simply initializing the weights of last fully connected layer (w_l) as orthonormal vectors and freezing this layer during training. While this maximizes interclass separation, this does not have any negative impact on in-domain classification accuracy.

2.3. OOD detection

Any test sample can be classified as OOD depending on whether it lies inside the region in feature space occupied by any of the known classes. Let us assume that OOD samples and samples from one of the known classes (l) are approximated by Gaussian distributions $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_l, \Sigma_l)$, respectively. The classification error probability between these two classes can be reduced by simply increasing their *Bhattacharyya distance* [13] that can be increased by making $\mathcal{N}(\mu_l, \Sigma_l)$ compact. The distribution of in-domain class l

Table 1. UCM data set split into in-domain, OOD for training, and OOD for testing sets. Among compared methods, only supervised DPN[forw] uses the OOD training classes. Proposed unsupervised method does not require them.

	Setting 1	Setting 2
In-Domain Classes	Beach, Chaparral, Overpass, Baseball Diamond, Intersection, Forest, Sparse Residential	Forest, Golf Course, Harbor, Tennis Court, Mobile Home Park, Freeway, Overpass
Out-of-Distribution Classes Training	Dense Residential, Freeway, Harbor, Storage Tank, Medium Residential, Runway, Tennis Court	Parking Lot, Sparse Residential, Chaparral, Buildings, Airplane, Storage Tanks, Agriculture
Out-of-Distribution Classes Testing	Agricultural, Airplane, Buildings, Golf Course, Mobile Home Park, Parking Lot, River	Medium Residential, River, Beach, Baseball Diamond, Runway, Dense Residential, Intersection

can be made compact by representing the class using its first singular vector $v_1^{(l)}$ of the training samples [11].

Given a test sample x'_n and its corresponding feature vector f'_n, an index can be computed as spectral discrepancy:

$$\theta_n = \min_l arccos(\frac{|f_n'^T v_1^{(l)}|}{||f'_n||}) \quad (2)$$

where θ_n is the minimum angular distance of the test feature vector from the first singular vector of any of the in-domain classes. Larger θ_n implies the test sample is distant from the in-domain classes and thus can be used as the uncertainty score to detect the OOD samples.

Table 2. Quantitative comparison for OOD detection. While supervised DPN[forw] obtains best result, unsupervised proposed method clearly outperforms the other unsupervised methods.

Model	**Setting-1**	**Setting-2**	**Supervision**
DPN[forw]	98.16	95.27	Supervised
Softmax	94.15	88.88	Unsupervised
ENN	86.32	90.53	Unsupervised
Proposed	97.67	94.40	Unsupervised

3. EXPERIMENTAL VALIDATION

3.1. Dataset and settings

We use the UC Merced (UCM) dataset [12] that contains images of 256×256 pixels size and comprises of 21 different classes, each having 100 samples. For evaluating the proposed method, we use the open-set recognition task where test set contains classes unseen during training. By dividing the 21 classes into three groups (in-domain, OOD training, OOD test) of 7 classes each (Table 1), we prepare the Setting-1. The in-domain dataset is further split randomly into 70% for training and 30% for testing. Similarly, we also form Setting-2 (Table 1). Our experiments use ResNet-50 architecture.

3.2. Compared methods

Following two methods are compared to the proposed method:

1. **Supervised DPN[forw]** [14] that uses OOD training data. While proposed unsupervised method cannot be expected to outperform this supervised method, it provides us an idea of how much the proposed unsupervised method lags behind the supervised methods.

2. **Unsupervised softmax** [9] based method that does not require any OOD training data.

3. **Unsupervised Evidential Neural Network (ENN)** [7] that does not require any OOD training data. Here we use the expected cross-entropy loss.

Performance is measured by Area under Receiver Operator Characteristic (AUROC). For compared methods, AUROC is computed based on maximum probability [5].

3.3. Result

The OOD detection performance of different methods is tabulated in Table 2. Performance is shown as $100\times$AUROC. For Setting-1, supervised DPN[forw] obtains best result. However, proposed method obtains similar result (only a difference of 0.49), in spite of being unsupervised. On the other hand, proposed method outperforms the other unsupervised methods by a large margin, i.e., ENN [7] by a difference of 11.35 and

softmax-based method [9] by a difference of 3.52. Similar result is obtained for Setting-2 where the proposed method outperforms unsupervised ENN by a margin of 3.87.

While improving the OOD detection performance, it is important to preserve the in-domain classification accuracy. For Setting-1, the proposed method obtains an in-domain classification accuracy of 99.21% in comparison 96.50%, 98.58%, and 98.75% obtained by ENN, softmax and DPN^{forw}, respectively. Thus, the proposed method actually slightly improves in-domain classification accuracy.

4. CONCLUSION

This paper proposed an unsupervised method for OOD detection. Unlike many other OOD detection method, the proposed approach is simple to implement and does not require significant modification of training loss functions. Our experiments on the UC Merced dataset demonstrated the effectiveness of the proposed method. While proposed unsupervised method cannot outperform the compared supervised method, it only lags behind by a little. On the other hand, proposed method outperforms the other unsupervised methods by a significant margin. Our future work will extend the method for OOD detection on different geographic areas and different sensors.

Acknowledgement

The work is funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

5. REFERENCES

[1] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[3] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, pp. 6402–6413, 2017.

[4] 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.

[5] Jay Nandy, Wynne Hsu, and Mong Li Lee, "Distributional shifts in automated diabetic retinopathy screening," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 255–259.

[6] Jakob Gawlikowski, Sudipan Saha, Anna Kruspe, and Xiao Xiang Zhu, "Towards out-of-distribution detection for remote sensing," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 8676–8679.

[7] Murat Sensoy, Lance Kaplan, and Melih Kandemir, "Evidential deep learning to quantify classification uncertainty," *arXiv preprint arXiv:1806.01768*, 2018.

[8] Shuyi Zhang, Chao Pan, Liyan Song, Xiaoyu Wu, Zheng Hu, Ke Pei, Peter Tino, and Xin Yao, "Label-assisted memory autoencoder for unsupervised out-of-distribution detection," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 795–810.

[9] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[10] Sudipan Saha, Biplab Banerjee, and Xiao Xiang Zhu, "Trusting small training dataset for supervised change detection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2031–2034.

[11] Alireza Zaeemzadeh, Niccolò Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah, "Out-of-distribution detection using union of 1-dimensional subspaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9452–9461.

[12] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010, pp. 270–279.

[13] Moataz MH El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Toward a tight upper bound for the error probability of the binary gaussian classification problem," *Pattern Recognition*, vol. 41, no. 6, pp. 2120–2132, 2008.

[14] Andrey Malinin and Mark Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 7047–7058.