THE INFLUENCE OF SAMPLING METHODS ON PIXEL-WISE HYPERSPECTRAL IMAGE CLASSIFICATION WITH 3D CONVOLUTIONAL NEURAL NETWORKS

Julius Lange[1], *Gabriele Cavallaro*[2], *Markus Götz*[2,3], *Ernir Erlingsson*[3], *Morris Riedel*[2,3]

[1] Humboldt University of Berlin, Germany
[2] Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany
[3] School of Engineering and Natural Sciences, University of Iceland, Iceland

ABSTRACT

Supervised image classification is one of the essential techniques for generating semantic maps from remotely sensed images. The lack of labeled ground truth datasets, due to the inherent time effort and cost involved in collecting training samples, has led to the practice of training and validating new classifiers within a single image. In line with that, the dominant approach for the division of the available ground truth into disjoint training and test sets is random sampling. This paper discusses the problems that arise when this strategy is adopted in conjunction with spectral-spatial and pixel-wise classifiers such as 3D Convolutional Neural Networks (3D CNN). It is shown that a random sampling scheme leads to a violation of the independence assumption and to the illusion that global knowledge is extracted from the training set. To tackle this issue, two improved sampling strategies based on the Density-Based Clustering Algorithm (DBSCAN) are proposed. They minimize the violation of the train and test samples independence assumption and thus ensure an honest estimation of the generalization capabilities of the classifier.

***Index Terms*—** Hyperspectral image classification, sampling strategies, clustering, DBSCAN, deep learning, Convolutional Neural Networks (CNNs)

1. INTRODUCTION

During the past few decades the processing of Earth observation data through remote sensing techniques has benefited from advancements in instruments on-board space and airborne platforms. Among all the possible products that can be derived from remote sensing data, classification maps are perhaps the most often used by many applications. Classification algorithms are utilized to distinguish between different types of land-cover classes in order to interpret processes, such as monitoring of urban growth, impacts of natural disasters, object detection, etc. When training samples are available, the model parameters of the classifier are learned in a supervised way. Once the training is completed, the main challenge is to obtain accurate and reliable semantic maps from previously unseen data. This capability is usually more influenced by the amount and quality of the training samples rather than the model complexity, since classifiers are based on the assumption that training and test samples are generated from the same feature space and distribution [1]. Remote sensing data usually present heterogeneous feature spaces and distributions due to differences in acquisition or changes in the nature of the object observed. As a consequence, most of the statistical models are likely to fail the prediction of new samples. A straightforward solution to this problem is to rebuilt from scratch the predictive model using new training data. However, these samples are usually either collected manually with ground surveys or automatically generated through image photo interpretation [2]. As a consequence there is a lack of appropriate benchmark datasets within the community and the practice of benchmarking new classification algorithms over a single image remains dominant.

Similarly to Liang *et al.* [3] and Hansch *et al.* [4] this paper aims at showing that the extraction of disjointed train and test sets though a random sampling approach cannot guarantee unbiased samples. However, this study considers two novel aspects. On the one hand, 3D CNNs are investigated as the spectral-spatial classifier. Due to the way convolutional neurons process a training sample within a receptive field, the overlap between the training and testing samples is artificially enhanced. One the other hand, to alleviate this overlapping effect, two alternative sampling strategies based on the DBSCAN [5] algorithm are proposed. The experiments, conducted on the full site hyperspectral Indian Pines dataset[1], confirm previous findings regarding random sampling techniques [3, 4] and show that the proposed sampling scheme leads to less biased error estimates. The amplification of the accuracy brought by the random sampling approach is attenuated, i.e., decreased for each class, while the performance evaluation can be considered fair, unbiased and with a rational estimation of the classifier generalization capabilities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 754304 DEEP-EST. The results of this research were achieved through the Human Brain Project PCP Pilot Systems at the Juelich Supercomputing Centre, which received co-funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 604102.

[1] https://purr.purdue.edu/publications/1947/1

2. BACKGROUND AND RELATED WORK

In order to reduce the need for and effort in recollecting training data, recent works have considered solutions based on transfer learning, domain adaptation and active learning approaches [6]. These solutions offer the capability of exploiting the knowledge acquired by the available ground reference samples for classifying new images acquired over heterogeneous geographical locations at diverse times with different sensors. However, Ball *et al.* [7] provide a summary of the common open-source hyperspectral datasets that are used for validating new deep learning classifiers methods. These entail four datasets, i.e., Indian Pines (small test site), Pavia University, Pavia City Center, and Salinas[2] and they are saturated in terms of classification accuracies.

The standard procedure for estimating the generalization error is to divide the ground truth samples into two disjoint sets, one for training and one for testing. The error obtained on the training data should not be considered since it is not difficult to decrease it to zero given a sufficiently complex method which can easily memorize the training data. Therefore, the sampling strategy that is adopted for producing these disjointed sets has a large influence for the validation phase. The random sampling strategy has been always considered as the natural choice, especially for classifiers that ignore the spatial information. Since spectral classifiers are less effective when dealing with very high spatial resolution images, modern classification pipelines include both spectral and spatial information. Recently, deep learning has brought in revolutionary achievements in many applications, including the processing of remote sensing images [7]. Remarkable results have been achieved with CNNs due to their hierarchical structure able to extract more hidden and deeper features. Recently, novel supervised CNNs have been proposed for hyperspectral image classification [8, 9]. These cover three-dimensional models that utilize receptive fields in both domains, spectral and spatial. The majority of these studies have carried out their experiments on standard hyperspectral datasets [7] by adopting random sampling strategies.

Researchers usually focus on improving the classification performance, while the above discussed problems are mostly neglected. The increase of spatially correlated data by spectral-spatial features and its influence on the quality of the estimate of the generalization error was already discussed by Zhou *et al.* [10]. A more recent work proposed a sampling scheme that minimizes the spatial overlap between train and test data [3]. The method aims to capture the full spectral variation of the image by globally sampling compact regions. Finally, Hansch *et al.* [4] evaluated different sampling approaches and proposed a new strategy that simulates a realistic gap of data variation between train and application phase. The method proposed in this work is a more flexible generalization of these two.

[2]https://goo.gl/QdLmUK

3. PROPOSED METHOD

3.1. Sampling Approaches

The idea of the proposed sampling approach is to minimize the number of biased samples. Bias occurs when directly neighboring or nearby pixels are present in both training and test sets. Due to their spatial closeness, information from one set may leak into the respective other, violating the independence assumption. In case of estimating central pixels based on a surrounding window mask, for example, spatial receptive fields in training and test data may overlap and be nearly identical. Correctly classifying a pixel of the same class in the test set based on the previously seen similar instance in the training data is very likely. In fact, the classification problem degrades from an actual pattern recognition to simple memorization. The proposed clustering-based method attempts to overcome this problem by, first, extracting larger contiguous regions using the class labels, e.g. buildings, fields, etc., and then distributing these disjointly between the training and test set. A bias, if present at all, would then only be relevant at the outer edges of such a region, but not for the inner pixels.

The extraction of the contiguous regions is achieved with the DBSCAN [5] clustering algorithm. It detects subgroups within a set through the recursive evaluation of a neighbor point density threshold ($minPoints$) criterion within a parametric search radius (ε) around a sample. Thereby, independent regions can be determined by clustering the coordinates of pixels of a particular class. Each resulting cluster directly corresponds to a region. The distribution of the identified regions between the training and test set is the next logical problem to address. In principal these regions could now be randomly sampled and assigned to either one of the two sets. However, the count of extracted regions is significantly lower (in the order of a few dozens) compared to the number of pixels. For this reasons, the likelihood of selecting an imbalanced training set rises strongly, e.g., one that does not contain patterns that are present in the test data, like cloud coverage for example.

Instead, an approach should be selected that maximizes the variability in the training set, so that a large number of potential patterns is covered. This requires to establish a metric that evaluates said variety. The first two, proposed as part of this work, are the region area size and statistical variance (σ^2). Based on this, sorting the regions in ascending, respectively descending order, and assigning them to the training set, up until the selected split percentage, should result in a less biased but highly variable pattern distribution. An example is depicted in Figure 1. Admittedly, employing the metric on all clustered regions before having splitting them into training and test data introduces bias itself. Namely, information from both, supposedly independent sets, is used to form them. Being from the same feature space and distribution [1], this means that the training set, as proposed, is treated favorably. Therefore, an overestimated out-of-sample accuracy on

(a) Random sampling. (b) Cluster sampling with area. (c) Cluster sampling with variance.

Fig. 1: Visualization of different sampling strategies exemplified using the class "forest" of the Indian Pines dataset. Black pixels are background, white training (10% of the available labeled samples) and purple test samples.

the test set should be the result, diminishing the true generalization capabilities. In the worst overestimation case, the prediction accuracy would be higher than a randomly sampled datasets. For practical applications, though, this bias is negligible as the experimental evaluation in Section 3.3 show.

3.2. 3D CNN and Dataset

The proposed 3D CNN is designed to perform pixel-wise classification of hyperspectral images. As input it accepts spatial-spectral tensors of size (w, w, c) (w window size; c number of spectral bands), exploits the spectral information and the correlation between neighboring pixels, and predicts the center pixel. The network is summarized by Table 1 and includes convolutional-, max-pooling-, fully-connected- and softmax layers. The triple alternation of convolutional and max pooling layers (i.e., applied to the c dimension) allows the network to reduce the number of channels and learn spectral features with different levels of abstraction. The output tensor of these layers is then flattened into a one-dimensional feature vector and passed to two fully connected layers for the class probability prediction. A softmax layer with a vector length corresponding to the total number of classes votes for the likeliest option. The experiments have been performed on the JURON pilot system at Jülich Supercomputing Centre and the development of the network was performed with the Keras library (2.0.8) and the TensorFlow (1.3.0) back-end.

Table 1: Complete set of specifications for the 3D CNN (with $583,962$ trainable parameters).

Feature	Representation / Value
Conv. Layer Filters	48, 32, 32
Conv. Layer Filter size	$(3,3,5)$, $(3,3,5)$, $(3,3,5)$
Pooling size	$(1,1,3)$, $(1,1,3)$, $(1,1,2)$
Dense Layer Neurons	128, 128
Activation Functions	rectified linear unit (ReLU)
Loss Function	mean-squared error (MSE)
Optimization	stochastic gradient descent (SGD)
Training Epochs	600
Batch Size	50
Learning Rate	1.0
Learning Rate Decay	5×10^{-6}

The dataset is the Indian Pines hyperspectral image acquired by the AVIRIS sensor in 1992 over an agricultural site composed of fields with regular geometry and with a variety of crops. It consists of 614×2166 pixels and 220 spectral bands, with a spatial resolution of 20m. The ground truth encompasses 58 different land-cover classes with a highly imbalanced density distribution. In the few works that considered this full size dataset [11], it has been a common practice to exclude the under-represented classes (e.g., with less than 100 samples) and discard noisy spectral bands. However, this work considers all the channels and classes in order to test the robustness of the classifier.

Table 2: Classification results (overall accuracy) of the sampling strategies with different % of training samples.

	10	30	60	90	%
Random	0.846	0.932	0.971	0.974	OA
	0.834	0.926	0.966	0.972	kappa
Area	0.289	0.323	0.381	0.615	OA
	0.231	0.245	0.318	0.581	kappa
Variance	0.251	0.334	0.358	0.389	OA
	0.207	0.267	0.285	0.322	kappa

3.3. Experimental Results

Most of the proposed CNN classifiers that considered the Indian Pines dataset as a benchmark used the small test site (16 classes in an area of 145×145 pixels) with random sampling strategies for dividing training and test samples. This leads to a vast number of classifiers that provided near perfect classification accuracy. When considering the full test site of Indian Pines, the state-of-the art classification accuracy (i.e., $\kappa = 0.84$ with 30% of the data to train with 20 classes and 20 channels excluded) was achieved by Romero *et al.* [11]. They proposed to use a greedy layer-wise unsupervised pre-training on deep CNNs coupled with a an algorithm for unsupervised learning of sparse features (i.e., Enforcing Lifetime and Population Sparsity - EPLS). The classification results achieved by the proposed 3D CNN are depicted by Table 2 and show that with only 10% of the available samples for training it was already possible to obtain $\kappa = 0.83$. These results confirm that

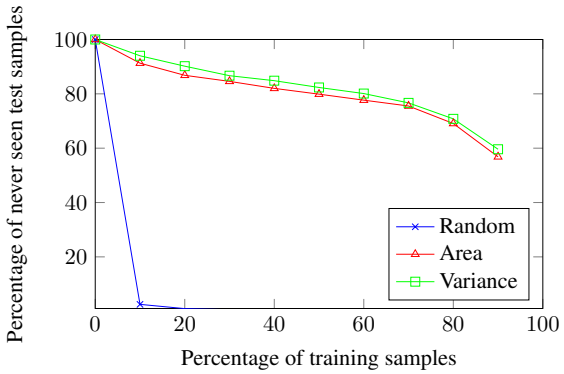

Fig. 2: Percentage of unbiased samples for the different sampling strategies with window tensor size w equal to nine.

random sampling approaches can always achieve the best results. However, the plot depicted by Figure 2 gives a clear explanation for these achievements. When considering the random strategy, the number of independent samples (i.e., not seen during the learning phase) are already less than 1% for a training set of 10%. On the contrary, the proposed sampling strategies allow to maintain an acceptable level of independence even for training set with higher amount of samples. On the one hand, this leads to worse classification results, as shown in Table 2. On the other hand, these numbers are a more trustworthy representation of how the resulting model could perform in real world applications, e.g., a usable transferable classifier. The classifier is unable to learn all the possible data variations like the same image acquired in different seasons. The gap of data variation between train and application phase remains in place.

4. CONCLUSIONS

The influence of different sampling strategies on the performance of pixel-wise image classification has been evaluated. Confirming previous research, the widely used random sampling approach violates the independence assumption due to the introduction of systematic bias. This is particular true for current state-of-the-art CNNs and the spatial overlaps in their receptive fields. The proposed sampling approaches using the DBSCAN clustering algorithm minimizes said bias and results in a classification accuracy on unseen test data closer to an actual out-of-sample performance.

In line with this observation, a more wide-spread adaptation of none-random sampling approaches for remote sensing classification problems stands to reason. Particularly, for transfer learning and concept drift problems the relevance of the presented findings is apparent. For the future, it is planned that other datasets are investigated using the proposed method. A similar classification accuracy performance degradation is to be expected. It will be of interest to further investigate the influence of the proposed and then added sorting metrics for the regions on the classifier performance.

References

[1] Q. Yang and X. Wu, "10 Challenging Problems in Data Mining," *International Journal of Information Technology and Decision Making*, vol. 05, no. 04, pp. 597–604, 2006.

[2] B. Demir, C. Persello, and L. Bruzzone, "Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, 2011.

[3] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the Sampling Strategy for Evaluation of Spectral-Spatial Methods in Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, 2017.

[4] R Hänsch, A Ley, and O Hellwich, "Correct and Still Wrong: The Relationship Between Sampling Strategies and the Estimation of the Generalization Error," in *Proceedings of the IEEE IGARSS*, 2017.

[5] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the SIGKDD*, 1996, pp. 226–231.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] J. E. Ball, D. T. Anderson, and C. S. Chan, "A Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools and Challenges for the Community," in *Proceedings of the SPIE Journal of Applied Remote Sensing*, 2017.

[8] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[9] M. He, B. Li, and H. Chen, "Multi-Scale 3D Deep Convolutional Neural Network for Hyperspectral Image Classification," in *Proceedings of the IEEE ICIP*, 2017.

[10] J. Zhou, J. Liang, Y. Qian, Y. Gao, and L. Tong, "On the Sampling Strategies for Evaluation of Joint Spectral-spatial Information based Classifiers," in *Proceedings of the 7th WHISPERS*, June 2015, pp. 1–4.

[11] A. Romero, C. Gatta, and G. Camps-valls, "Unsupervised Deep Feature Extraction for Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1–14, 2015.