DEEP CONVOLUTIONAL NETWORKS WITH SUPERPIXEL SEGMENTATION FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Jiayan Cao[1, 2, 3]*, Zhao Chen*[1, 2, 3]*, and Bin Wang*[1, 2, 3*]

1. Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University,
Shanghai 200433, China;
2. State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University,
Beijing 100875, China;
3. Research Center of Smart Networks and Systems, School of Information Science and Technology,
Fudan University, Shanghai 200433, China

ABSTRACT

To combat the well-known Hughes phenomenon occurred in hyperspectral classification, most of the previous works adopt dimensionality reduction or manifold learning technique before supervised learning. While in this paper, we propose a different scheme: First, we design a pixel-wise classifier based on Convolutional Neural Network that could directly mapping observed spectrum to class distribution. Then, we conduct superpixel segmentation on the prediction map that learned by previous model and output the final classification results by spatial and spectral factors jointly. Varied from other deep learning method, our classification framework learns and infers spectrum efficiently via deep hierarchy with convolutional and pooling layers, thus forming a direct relationship between high-order data and class distribution. Moreover, superpixel segmentation helps further boost the accuracy of the classification by combining the spatial information. In experimental studies, multiple hyperspectral datasets with various context and spatial resolution are used to validate the proposed method. The experimental results show that the proposed method is efficient and competitive in practical uses.

Index Terms— Hyperspectral Image Classification, Deep Learning, Convolutional Neural Network, Superpixel Segmentation, Support Vector Machine

1. INTRODUCTION

The accurate classification on Hyperspectral imagery (HSI) is of great importance to many of the undergoing application including agriculture, surveillance, and astronomy, etc. Unlike human visual system, HSI divides spectrum into considerable amount of bands and this plentiful information on spectrum helps analyzing underlying materials and identifying desired objects. However, with the overwhelming growth of spectral channels, "curse of dimensionality" significantly affects the generalization ability of supervised classifiers, thus leading to disastrous prediction results.

During the past decade, various deep learning frameworks have been implemented on numerous topics and domains of machine learning. An interesting aspect of deep learning is that instead of requiring prior feature extraction process, it learns relationship between input and output directly from enormous amount of data, showing great flexibility and capability than traditional shallow networks. However, given limited training samples, deep learning model suffers from complicated architecture and behave poorly on validation sets.

Consequently, we proposed a deep learning based classification architecture that will learn and generalize well not only under small sample conditions but also given plenty of data: First, Convolutional Neural Network (CNN) is utilized and fine-tuned to directly map observed spectrum to corresponding class distribution for each pixel; Then, superpixel segmentation is performed on the prediction map that learned by CNN; Final classification result is output via collaboratively fusing class distribution predicted by CNN and spatial correlation extracted from segmentation step. From the perspective of Bayesian learning, proposed architecture incorporates superpixel segmentation into vanilla CNN structure where CNN learns and predicts observed spectrum pixel-wise of maximum likelihood and segmentation part serves as a local prior which further promotes the efficiency of supervised learning.

2. RELATED WORKS

Traditional approaches [1] treat each pixel separately and classify only by spectrum signatures. Before feeding raw input to the classifiers, researchers tend to apply feature extraction, dimensional reduction and band selection to

This work was supported in part by the National Natural Science Foundation of China under Grant 61572133 and in part by the Research Fund for the State Key Laboratory of Earth Surface Processes and Resource Ecology under Grant 2015-KF-01.

tackle with the dilemma between limited samples and overfitting. A competent method, support vector machine (SVM), was introduced into classification applications [2], which later proved to be a baseline method for classification.

In terms of neural nets, traditional models including logistic regression and SVMs are shallow feedforward networks while machine learning systems with multiple layers are powerful enough to extract more abstract and invariant features, thus contributing to higher accuracies than shallow models. More importantly, multiclass SVM is a generalization of the basic 2-class SVM with either 1-versus-all or 1-versus-1 strategy. Predicted label is created in form of hard classifiers without giving entire information on class distribution, thus obscure in Bayesian inference.

In literature [3], Stacked Auto-Encoder (SAE) based deep learning framework is first applied into hyperspectral classification and has already achieved promising results among state-of-the-art methods. It requires huge epoch time for pre-training and fine-tuning phases to keep a low reconstruction and generalization error. As a consequence, more than 60% of total labeling samples are used to train a well-tuned classifier of such enormous fan-in and fan-out capability. However, in real-world implementation, limited samples are provided rather than huge amount of efforts of labeling.

3. PROPOSED CLASSIFICATION ARCHITECTURE

To fully understand the proposed classification architecture, the inner structures and overall workflow of joint spatial-spectral classification framework is elaborated as follows.

3.1. Convolutional Neural Network

The essential functionality of CNNs [4] is to extract simple features at higher resolution and then convert them into more complex feature at coarser scale. The cascading convolutional and pooling layers behave as trainable feature detectors while the output layers are fully optimized and dedicated for classification tasks. Block diagram for a vanilla CNN is shown in Fig. 1.

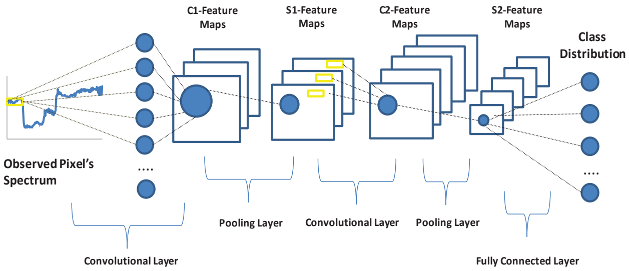


Fig. 1. Block diagram for a vanilla CNN applied in hyperspectral classification

In hyperspectral imaging scenarios, we select and vectorize pixel by spectrum from the raw HSI and classify it via spectral features. The low levels are composed to alternatively cascading between convolutional and pooling layers. Different from implementation in other computer vision tasks, we design all the convolutional filters in CNN to be 1D rather than 2D, which could learn the correlation between various spectrums. We gradually increase the number of output feature maps as the layer goes deeper and preserve as much information as possible in the pooling layers. The upper levels are fully connected and correspond to a traditional multi-layer perceptron structure. Labels for training sets are one-hot coded. Thus the prediction of class for test set is normalized by non-linear functions (sigmoid, softmax and etc.), forming a single layer logistic regression.

3.2. Superpixel Segmentation

Due to the physical principle on radiation interaction, pixels in HSI data are largely mixed with multiple endmembers. Prior knowledge on each class for the current pixel is often consistent with those of its spatial neighbors in local areas. As a result, after supervised learning, grouping pixels into local clusters is a useful step in analysis of classification.

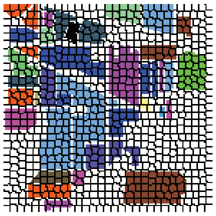


Fig. 2. Visualization of SLIC superpixels on Indian Pines, boundaries colored in black

In our proposed model, we implement Simple Linear Iterative Clustering (SLIC) [5] to extract local cluster through both spectral and spatial similarity measurement. SLIC superpixel can be computed in the following manners:

1. Initializing set of cluster centers by sampling pixels at step size S. Move cluster center to the lowest gradient position in a 3×3 neighborhood
2. Assign each pixel to the closest cluster center in a $2S\times 2S$ region according to the unified distance:

$$d=\sqrt{\frac{\left\|x_{i,j}-x_{p,q}\right\|^2}{m}+\frac{\left\|c_{i,j}-c_{p,q}\right\|^2}{S}} \quad (1)$$

   Where $x_{i,j}$ and $c_{i,j}$ represents spectral signature and spatial coordinates of current pixel. m controls the relative importance between spatial and spectral similarity. In experiment, we set $m=10, S=5$
3. Update each cluster center.
4. Repeat clustering until some threshold.
5. Connect disjoint segments to be connected to largest neighboring clusters.

To illustrate the effects of SLIC, extracted superpixels on Indian Pines and Pavia University are visualized in Fig. 2. Clusters within the same class preserve local similarity among neighboring pixels and adhere well to spatial boundary.

3.3. Joint Spatial-Spectral Classification

From perspective of Bayesian learning, classification problem is equivalent to an estimation on maximum a posteriori probability:

$$p(c|x) \propto p(x|c)\,p(c) \ , \qquad (2)$$

where $p(x|c)$ is the likelihood probability approximated by supervised learning and $p(c)$ is the prior on each class. Regularizations on the class prior probability are assumed to remedy overfitting problem that may occur in supervised training. Accordingly, CNN is fine-tuned to estimate test samples in maximum likelihood criteria and SLIC superpixels provide a local prior for spatial smoothness.

The overall workflow for the proposed classification framework is shown below in Fig. 3. At first, CNN is trained by labeled samples. On the output layer, class prediction is computed for each pixel in the HSI, obtaining maximum likelihood estimation on test samples.

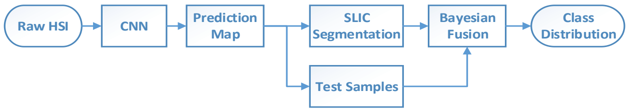


Fig. 3. Overall workflow of the proposed joint spatial-spectral classification framework

Next, superpixels are generated by the prediction map rather than raw HSI data. The reduction of bands on input dimensions helps facilitate the execution time without losing much of the accuracy. After gaining superpixel centroids, we add up counts of training samples and predicted labels that fall into the corresponding cluster on each class and normalize it to be a local prior probability. In the final step, class distribution for each pixel is calculated by multiplying pixel-wise likelihood with prior probability in the nearest cluster and the computed result is maximized to ensure a robust estimate on both spectral and spatial features.

Compared to SAE based deep networks, CNN has much more compact structures than SAE and learns drastically varying spectral signatures from the raw data. Furthermore, with the help of superpixel segmentation, the proposed model is able to deal with small sample problem that often deteriorates the overall classification accuracy of supervised learning.

4. EXPERIMENTAL RESULTS

Before proceeding into the following experiments, the architecture of CNNs, which serves as a core module in the overall workflow, is carefully designed. Then, we compare convolutional nets with SVM-based approaches.

Previous works tend to fine-tune deep neural nets on enormous amount of training samples, while in our experimental settings, we try to figure out how our proposed framework will behave and perform given limited amount of training proportion (TP).

4.1. Dataset Description

Three hyperspectral datasets with different context and spatial resolution are used to validate our proposed methods.

1. Indian Pines: This dataset was captured in 1992 by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in Northwestern Indiana, a mixed agricultural/forest area[1]. The image consists of 145×145 pixels and 220 spectral channels in the wavelength range from 0.4 to 2.5 μm . Before experiments, bands 104-108, 150-163 and 220 have been removed due to noise or water absorption phenomena.
2. Kennedy Space Center (KSC): AVIRIS acquires data in 224 bands of 10 nm width with center wavelengths from 400-2500 nm over the Kennedy Space Center, Florida, on March 23, 1996. The KSC data have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands were used for the analysis and 13 classes representing the various land cover types that occur in this environment were defined for the site.
3. Pavia University (Pavia U): Pavia Univ. dataset is acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The number of spectral bands is 103 for Pavia University. Spatial resolution is 610×610 pixels and ground truth differentiates 9 classes each.

Ground truth maps and classification results of 3 datasets are illustrated respectively in Fig. 4, Fig. 5 and Fig. 6. In order to quantitatively compare the performance of various classifiers, overall accuracy (OA), average accuracy (AA) and Kappa coefficient (κ) are used as measurement indices. Every numerical result is the average of 10 runs, each one using a set of randomly selected training samples.

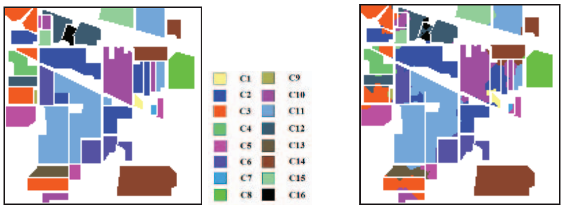
4.2. Architecture of CNNs

The deep convolutional network consists of 2 subsequent layers. Each layer is composed of cascading structure with convolution and polling stages. Pooling stages are all set to be subsampling by a small factor of 2. For the convolutional

[1] http://dynamo.ecn.purdue.edu/biehl/MultiSpec

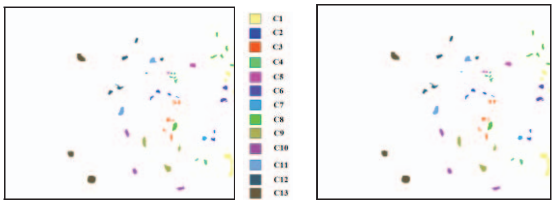
layers, we set the kernel size to 5 and number of feature maps to 6 and 12 respectively for 2 convolutional layers.

Mini-batch strategy is adopted to update trainable parameters in the nets and size of the training batch is set to 10 samples each. Learning rate is controlled as 0.5 and number of training epochs is set to 400 to ensure the nets converges both quickly and accurately.



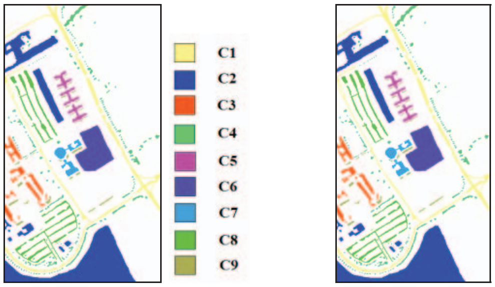
(a) (b)

Fig. 4. Indian Pines (a) Ground truth map (b) Examples of the classification results by SLIC-CNN (TP=5%, OA=94.17%)



(a) (b)

Fig. 5. KSC, (a) Ground truth map, (b) Examples of the classification results by SLIC-CNN (TP =5%, OA=99.17%).



(a) (b)

Fig. 6. Pavia Univ., (a) Ground truth map, (b) Examples of the classification results by SLIC-CNN (TP =5%, OA=96.77%).

4.3. Comparison with Other Approaches

In this section, we mainly compare the behaviors of CNN with RBF-SVM based classifiers. In joint-spatial learning, we consider the performance gap between SLIC-CNN and RBF-SVM with spatially fused input (SRBF-SVM) using window sliding techniques (window size is 5×5). As mentioned before, training proportion is either 5% or 30% of total labeled samples to simulate conditions under limited and sufficient number of labeled samples.

Table 1 gives detailed information on classification results. CNN proves to be a competent pixelwise classifier with RBF-SVM. With the help of SLIC segmentation, the overall accuracy is further boosted and outperforms other SVM based approaches. In small sample settings, the proposed classifier achieves satisfying result among other algorithms.

Table 1. Comparison of CNNs with other algorithms

Dataset	TP (%)	Criterion	RBF-SVM	SRBF-SVM	vCNN	SLIC-CNN
KSC	5	OA	84.41	90.73	85.08	**99.17**
		AA	77.72	84.62	78.69	**98.97**
		κ	0.8262	0.8965	0.8338	**0.9908**
	30	OA	92.12	95.67	92.53	**100.0**
		AA	87.71	91.91	88.30	**100.0**
		κ	0.9122	0.9517	0.9168	**1.000**
Pavia Univ.	5	OA	87.46	90.66	89.05	**96.77**
		AA	82.45	82.14	84.52	**92.88**
		κ	0.8297	0.8753	0.8540	**0.9571**
	30	OA	93.89	95.19	92.93	**99.64**
		AA	91.68	91.34	91.16	**99.13**
		κ	0.9185	0.9361	0.9062	**0.9952**
Indian Pines	5	OA	59.53	80.84	71.51	**94.11**
		AA	37.15	64.96	61.23	**86.97**
		κ	0.5144	0.7783	0.6730	**0.9329**
	30	OA	76.65	95.45	84.53	**97.24**
		AA	58.02	90.70	82.11	**95.13**
		κ	0.7293	0.9480	0.8232	**0.9685**

5. CONCLUSION

In this paper, we combine superpixel segmentation and deep convolutional neural network to create a novel classification framework on hyperspectral classification. The proposed method involves fine-tuned CNNs that learn and infer directly from the raw data and adaptive mechanisms of spatial superpixel clustering that further boost the accuracy of undergoing tasks. Compared to other deep learning based methods, we design and test our deep nets under limited samples conditions rather than big training data, gaining promising classification results. Nevertheless, SVM-based classifiers discard redundant information on class prediction while our proposed method utilizes both likelihoods on spectral and local spatial features as a robust estimation in the sense of Bayesian inference.

6. REFERENCES

[1] Campsvalls G, Tuia D, and Bruzzone L, “Advances in hyperspectral image classification.” *IEEE Signal Processing Magazine*, 31.1, pp. 45-54, 2014

[2] Melgani F., and Bruzzone L., “Classification of hyperspectral remote sensing images with support vector machines.” *IEEE Trans. Geosci. Remote Sens.*, 42(8), pp. 1778-1790, 2004

[3] Chen Y., Lin Z., Zhao X., Wang G., and Gu Y., “Deep Learning-Based Classification of Hyperspectral Data.” *IEEE J-STARS*, 7(6), pp. 2094-2107, 2014

[4] Lecun Y., Bottou L., Bengio Y., and Haffner P., “Gradient-based learning applied to document recognition.” *Proceedings of the IEEE*, 86(11), pp. 2278-2324, 1998

[5] Achanta R., Shaji A., and Smith K., “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods.” *IEEE Trans. PAMI,* 34(11), pp. 2274-2282, 2012