LEARNING DEEP SPECTRAL FEATURES FOR HYPERSPECTRAL DATA USING CONVOLUTION OVER SPECTRAL SIGNATURE SHAPE

Shailesh Deshpande[1], *Rohit Thakur*[2], *Balamuralidhar P.*[3]

[1, 3]Embedded System and Robotics, TCS Research, India
[2]Analytics and Insights, TCS, India
{shailesh.deshpande@tcs.com}

ABSTRACT

Deep convolutional neural networks learn the spatial image features automatically, for classifying a hyperspectral image. Learning the spectral features automatically is equally important in analyzing the hyperspectral image. However, most of the earlier work treat a hyperspectral pixel as a *n* dimensional vector (*n* = no. of bands) and a separate convolution is performed over the depth. The features so learned are stacked together with the spatial features and are used for further processing. The semantics of the learned spectral features are completely ignored and are not interpretable in these approaches. We propose a simple transformation of the hyperspectral pixel to two-dimensional spectral graph (shape) and then the convolution over the same. This results in learning the spectral features that can be interpreted using spectroscopic knowledge of the material. We compared our approach with some of the common deep learning approaches for the hyperspectral data. The improvements are evident from the experiments.

Index terms – Hyperspectral signature, deep features, Convolutional Neural Network (CNN), spectral features, Capsule network

1. INTRODUCTION

One of the advantages of a convolutional neural network (CNN) is that it learns the features and classification model simultaneously. In the case of image classification or segmentation, the learned features are spatial features (like edges, arcs, convex shapes-contours of objects and object parts etc.) and are interpretable. There is an inherent hierarchy present in these features. The feature maps at the lower (closer to input) layers show the primitive image features such as edges with different orientations. The feature maps at the different higher levels show the higher-level features such as arcs, shapes, and object parts respectively. Thus, the features show the parts or whole of the primal sketch progressively and are consistent with the visual semantics. Often, the spatial dimensions (row and columns) of the image are convolved to build the hierarchical features. The convolution over the spectral domain is not performed for a visible or a multispectral image. 3-D kernel volume (say 3x3x*n=number of bands*) with the depth equal to the number of bands of the image is commonly used. This works for the visible or the multispectral image as the spectral features are not dominant for them. However, the spectral information is critical for classification of the hyperspectral (HS) image. There is increasing interest in using CNNs for the hyperspectral image (HS) classification. However, the common architectures for the visible or multispectral images that perform convolution over the spatial dimension are not useful in the classification of HS image. The recent works on CNN application to the HS image have recognized the drawbacks and attempt incorporating the convolution over the spectral dimension as well.

The early work by Zabalza et al. [1] used a two layered architecture to reduce the dimensionality of the spectral signature. They represented a segment (of the spectrum, a wavelength range) of the spectral signature by an autoencoder and then created the complete spectrum by combining the segment-wise representations. The idea was to understand importance of the few segments, which were selected using covariance matrix. Dai et al. [2] represented a pixel spectrum as a row or column vector and it was found useful in understanding the deep features in waveform. Their architecture did not use 4096 dimensional fully connected layer. Instead, an average pooling layer was used. This representation of a hyperspectral pixel is common in the literature.

Zhao et al. [3] used, in their related work, different dimensionality reduction techniques to reduce the dimensions of the HS data. The deep spatial feature for a HS pixel were learned using a small window around the pixels. The deep features and the reduced HS features were stacked together for further classification. In these approaches the spectral features (bands) were selected using standard dimensionality reduction techniques and no neural transformation was performed over an input spectrum. Instead of using dimensionality reduction methods such as (Principal Component Analysis) PCA, Abdi et al. [4] represented the spectrum using deep autoencoders. They were stacked together with the PCA spatial features from the window around the pixel and processed further.

Mou et al. [5] treated the spectrum of a pixel as a sequence of correlated values. They used a Recurrent Neural Network (RNN) to label the pixel. Each recurrent unit of the network was trained using inputs from a particular band in the sequence (grey values of the spectrum for each band) and the output from the previous network. Thus, it attempted to capture the spectral relation between grey values of the two consecutive bands in the spectrum. The proposed network by Gao et al. [6] learned a set of features using Attribute Profiles (probably from a grey image of each band). Then, each feature was processed by a CNN block parallelly. Finally, the features learned by the CNN were stacked together to learn the labels (of the pixel) using the Softmax.

Arun et al. [7] used an approach similar to the [3, 4]. They used the stacked feature vectors from the spectral and the spatial features learned using convolutional network. The capsule network (CapsuleNet) [8] was used to process the stacked feature vectors. In the alternative implementation, the spectral features were learned using CapsuleNet. However, the capsules which are supposed to identify the primitive features (and its hierarchy of other features in spectral domain) did not perform that task. The authors [7] do not explain the role of the capsules played in detecting the spectral features/primitive, if any. The elaborate discussion is required to explain the role of the capsules played (if any) in building the spectral feature hierarchy as the CapsuleNet is designed to learn by dynamic routing.

As can be seen in the literature, the convolution is performed over a n-dimensional pixel vector where n = number of the bands. The pixel vector is stacked vertically or horizontally, and the convolution is performed over the length of the vector. Additionally, the spatial features are stacked up together with the spectral features and the joined vectors are processed further. Thus, it is supposed to learn the deep features in the spectral domain. However, the features extracted by the spectral convolution (in this manner) are difficult to interpret in the spectral domain. For example, the spatial convolution learns the features that represent edges, arcs, shapes etc. (contours of the object part/objects) and similarly, the spectral features learned by the spectral convolution should not be just specific wavelengths/bands, but diagnostic absorptions or reflections as of the spectral signature at a given position, slope or convexity or concavity of the spectral shape and so on. None of the approaches used in the past literature (barring a few) attempt addressing this problem. They do not focus on the spectral features and its interpretation like the interpretation of the spatial features in spatial domain. The convolution over a one-dimensional spectral vector is difficult to interpret as it appears to learn discontinuities in the pixel vector for a given receptive field. Visualizing the same is, again, a challenging task as the vector is one dimensional. Building the hierarchy of these spectral features like the spatial features is difficult. Furthermore, the results are shown on the dataset like Pavia [9], which is having very high spatial resolution. Most of the accuracy improvement for such a high-resolution dataset is because of the spatial features. Whether the vector based spectral convolution is useful in medium to coarse resolution imagery needs to be investigated further (as the spatial features are not dominant as compared to spectral features, in the images). The systemic mechanism of the spectral and the spatial features learning for the imagery with the dominant spectral features such as HS is required.

In contrast to that, the alternative mechanisms that can extract semantically consistent/interpretable spectral features are a) a one-dimensional convolution by kernels dedicated for the different wavelength segments, possibly each one without sharing the weights. For example, the kernel extended over the entire range of the spectrum (or fully connected layers with neurons equal to number of bands) in first convolutional layer (a single kernel) would learn significance of the different wavelengths and so on; b) a two-dimensional convolution over the spectral shape, that is, a graph of the spectrum in XY space, or a line diagram of the spectral signature. The latter is more intuitive for multiple reasons. When the expert analyzes the spectrum of a given material, he observes the shape of the spectrum at various wavelengths, that is, the diagnostic absorption, the convexity of the curve, the slope at different positions or over a particular wavelength range etc. Furthermore, when there are no specific diagnostic features for the given materials, the expert comprehends and compares the entire spectrum shape for discriminating the materials. The shape of the parts and/or the complete spectrum are intuitively useful. The shape features such as arcs and arc segments formed by joining multiple arcs show a hierarchy of the features as well. Thus, the representation would further aid in architecting the robust capsule network [8].

In this paper we focus on approach b; we transform a HS pixel vector into the two-dimensional spectral shape and then perform the convolution over the image of graph thus formed. The intuition is to learn the spectral features as represented by the shape of a spectrum or in other words the features which a spectroscopy expert uses to interpret the spectrum. Thus, now the filters would learn edges, arcs, arcs segments and the other shape features of the spectrum as applicable. We trained our architecture using cross entropy loss. Some of the elements of the proposed architecture are inspired by the similar architectures in the literature. In our preliminary experiments we compared our results with the standard pixel vector representation and the capsule network. We believe as the spectral features hierarchy is accessible, this representation of the pixel vector would be more suitable to CapsuleNets. We use only spectral features in the current experiments and classification is performed at the pixel level without any spatial contextual information.

2. PROPOSED METHOD

2.1 Feature representation

We plot the graph of a n dimension spectral signature and use the image of that graph/line chart as an input for our CNN architecture. As the spectral signature is converted to the shape, we can decompose this shape using hierarchical features learned at the different convolution layers at different levels. This is like recognizing a handwritten digit using its image. Now, because of this transformation, the deep architecture would be able to learn the features that are consistent with the spectroscopic interpretation. The features learned by deep architectures would be similar to spectral signature features as domain expert seek identifying the material corresponding to the signature. We created 128x128 image from the pixel vector. The choice of the image size, at present, is determined by the number of bands in the dataset. The parameter is configurable and can be changed according to the number of bands in the working dataset. Any size that represents the graph without loss of information is suitable.

2.2 Architecture

We propose two architectures; one inspired by the one dimensional waveform architecture [2] and one by the CapsuleNet [8]. We describe the one dimensional architecture first.

The architecture comprises of the three blocks of convolution followed by the one block of fully connected layer. Each block of the convolution consists of the three convolution layers and each convolution layer is followed by a batch normalization layer. The last convolution layer of each block is also followed by a dropout layer to prevent the over-fitting. We maintain the drop-out rate of 40%. The first two convolution layers in each of the blocks are having the kernel size of 3x3 and the last convolution layer of each block has the kernel size of 5x5. The kernels use strides of three and five respectively. We have the two convolution layers of 3x3 instead of using one convolution layer of 5x5 as we can mimic 5x5 by using two consecutive 3x3 layers and it will be more nonlinear. We do not use max pooling in the network as important information is lost in the process of max pooling. Instead, we use the convolution layer with strides three and five for sub-sampling. Each convolution layer of all the convolution blocks consists of 32, 64, 128 feature maps respectively. The fully connected block of the network consists of the two fully connected layers. The first fully connected dense layer consists of 256 units whereas the second fully connected layer depending upon the number of classes considered has nine units in it (Figure 1).

The second proposed architecture is the CapsuleNet modified with the reconstruction loss. This is designed to further enhance the performance and study loss functions for the spectral classification. Rest of the network is similar to the CaspsuleNet [8].

We have used Rectified Linear Unit (RELU) activation function for all the layers in the network. We have used the Softmax activation in the last layer. We have performed one hot encoding over the label so that we can calculate the loss between the predicted and actual value by using the categorical cross entropy loss function. Categorical cross-entropy compares the prediction distribution with the true distribution, where the true class is represented as a one-hot encoded vector, and the closer the predictions are to that vector, the lower the loss. We have used Stochastic Gradient Descent (SGD) with learning rate of 0.01 and momentum of 0.5 as SGD with momentum tends to reach better optima and has better generalization than adaptive optimizers [10].

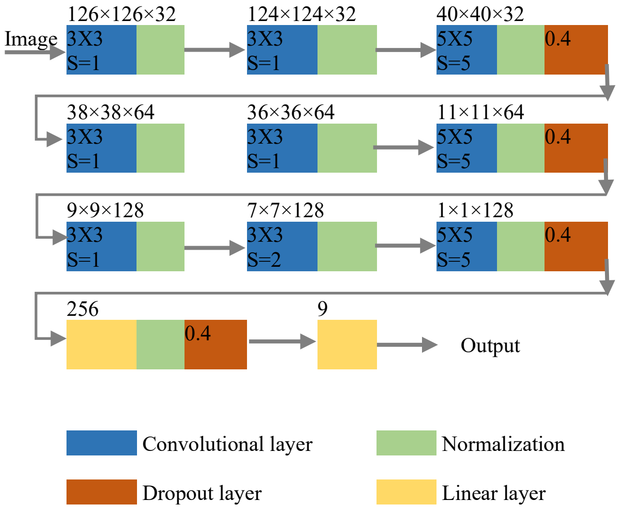

Figure 1. Block diagram for architectures P1 and P2 (see section 2.3)

2.4 Experimentation details

The dataset used was the well-known Pavia university dataset of the hyperspectral image. We have used Pavia university image. It has 103 bands covering 430 nm to 860 nm wavelength range and has spatial resolution of 1.3 m. The image contains the nine classes namely asphalt, meadows, gravel, trees, painted metal sheets, bare soil, bitumen, self-blocking bricks, and shadows. As we wanted to explore the semantics of the spectral deep features, we used only spectral features of the pixels for classification, by ignoring the spatial information of the pixel/s completely.

We performed multiple classification experiments using the proposed architectures and the other standard architectures reported in the literature. 1-d-wav is the architecture as implemented by [2]. P1 is our architecture as described in section 2.3. The architecture P2 is P1 with the reconstruction loss (Figure 1), the P3 is the standard CapsuleNet [8], and the P4 is the CaspsuleNet with reconstruction loss (Table 1). For each of the architectures, the pixel vector and the spectral shape as proposed were provide as the input for comparative assessment. We performed five-fold cross validation deploying stratified split to assure that the proportions of the classes in the training and the validation data sets are maintained.

3. RESULTS AND DISCUSSION

3.1 Accuracy improvements over standard representation

The accuracy improvements for the spectral shape input are evident from table 1 (and 2). The transformed pixel vector provided added advantage over the pixel vector in all the architectures. The improvement is substantial for all the architectures. Furthermore, if we compare the traditional layered architecture with the CapsuleNets for the spectral features, the CapsuleNets performance is enhanced. This enhancement can clearly be attributed to semantically consistent spectral features learned. Furthermore, preliminary experiments on loss functions suggest that the two-way optimization, that is, the Softmax loss and the reconstruction loss help in improving accuracy. The computational efficiency is the added advantage as no spatial features for the pixel were considered.

Table 1. Comparison of classification accuracies (in % age)

	Pixel Vector	Spectrum Shape
1d-wav	41.00	NA
P1	45.00	79.56
P2	46.73	81.24
P3 cap	32.81	84.63
P4 cap	52.18	91.71

Table 2. Precision and recall for CapsuleNet (P4) for one of the folds for spectrum shape input (in % age)

Class	Precision	Recall	F1-score	Support
Asphalt	**96.6**	**91.0**	**93.7**	**1421**
Bare soil	**82.8**	**98.0**	**89.8**	**1030**
Bitumen	**79.5**	**90.6**	**84.7**	**287**
Gavel	**82.1**	**87.0**	**84.5**	**432**
Meadows	**99.5**	**92.2**	**95.7**	**3951**
Painted-metal	**96.6**	**100**	**98.3**	**284**
Bricks	**85.5**	**89.8**	**87.6**	**729**
Shadows	**92.8**	**100**	**96.2**	**218**
Trees	**90.6**	**98.1**	**94.2**	**648**
Accuracy			**93.0**	**9000**
Weighted avg.	**93.6**	**93.0**	**93.2**	**9000**

3.2 Visualization and interpretability of the spectral features

We visualized the learned spectral features by pictorially representing the activations by various layers of proposed architectures (Figure 2). The proposed transformation of the HS pixel vector to the two-dimensional spectral shape as an input to the deep neural network enable learning of spectral features that are semantically interpretable. Deep architecture designed for processing this transformed signature learns the spectral features automatically. These features are the same as the spectroscopic experts seek in the material for its identification. The spectral features that are learned in different layers of the convolution reflected the hierarchy of spectral features as well. For example, some of the short arcs of spectral shape learned are common diagnostic features of the signatures (Figure 2). Further, if the shape of the signature over a specific wavelength rage is important, it is activated appropriately. Furthermore, the lower level primitives can be successively combined to form the spectral signature (just like spatial convolution creates a hierarchy of spatial features).

3.3 Interpretability of Capsules

The CapsuleNets are specifically designed to identify primitives and build semantically coherent association between the primitives found in the two convolution layers. That is, the parts of the objects and objects are associated using dynamic routing. This behavior is not spectrally reproduced by the CapsuleNets reported earlier as they use the pixel vector in deep processing. However, we can build the semantically sensible hierarchy of the spectral features because of the transformation of the one-dimensional vector to the two-dimensional spectral signature (Figure 2). These preliminary experiments indicate potential of the shape features for extracting the spectral feature hierarchy using the capsule nets.

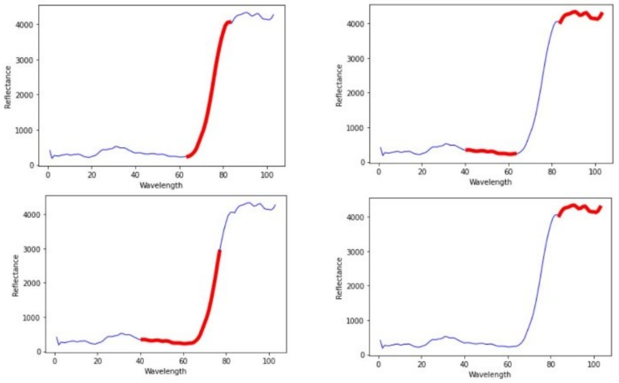

Figure 2. Visualization of the spectral features learned by the capsules. Each individual graph shows activated part of the spectral shape, in red, by some of the first-layer capsules. From top to bottom and left to right; shows red edge, red absorptions and infrared reflectance, red absorption, and infrared reflection as activated features respectively. These are the well-known spectral features for vegetation. The activations indicate that the capsules learn the spectroscopic features for vegetation when spectral shape is provided as the input, which is not possible with the one-dimensional pixel vector. Please note that the actual activations are enhanced pictorially for illustration purpose.

4. CONCLUSION

Most of the CNNs for classification of a HS image use *n* dimensional pixel vector (where *n* = number bands) for spectral convolution. However, the spectral features learned by convolution over *n* dimensional pixel vector are not easily interpretable in spectroscopic domain. We proposed two alternatives for extracting the spectral features that are consistent with spectroscopic interpretation. In the present work we transform the pixel vector to a spectral shape and then provide that as an input to the CNN. The preliminary results indicate the potential of the spectral shape as the input instead of the pixel vector. Further investigation is required to understand the robustness of the approach. The alternative image representations and the convolution approaches need to be investigated further.

In addition, the effect of the loss function on the classification accuracy is interesting topic. The triplet loss may provide further enhancement over the reconstruction loss. The formulation can be very similar to finding a match for a given photograph in the database of similar and dissimilar candidates. The triplet loss is especially desired as the spectral signatures of many urban materials are similar to each other. In this case, minimizing the distance between the similar pairs and maximizing the distance between the dissimilar pairs of the signatures would be very helpful.

5. REFERENCES

[1] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du and S. Marshall, "Novel segmented stacked auto encoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," Neurocomputing, vol. 185, pp. 1-10, 12 April 2016.

[2] W. Dai, C. Dai, S. Qu, J. Li and S. Das, "Very deep convolutional neural networks for raw waveforms," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2017.

[3] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," IEEE Transactions on Geosciencce and Remote Sensing, vol. 54, no. 8, 2016.

[4] G. Abdi, F. Samadzadegan and P. Reinartz, "Spectral–spatial feature learning for hyperspectral imagery classification using deep stacked sparse autoencoder," Journal of Applied Remote Sensing, vol. 11, no. 4, 2017.

[5] L. Mou, P. Ghamisi and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3639-3655, 2017.

[6] Q. Gao, S. Lim and X. Jia, "Hyperspectral image classification using convolutional neural networks and multiple feature learning," Remote Sensing, vol. 10, no. 2(299), 2018.

[7] P. Arun, K. Buddhiraju and A. Porwal, "Capsulenet-Based spatial–spectral classifier for hyperspectral images," Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 6, 2019.

[8] S. Sabour, N. Frosst and G. E. Hinton, "Dynamic Routing Between Capsules", 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017.

[9] B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, H. van der Piepen, and M. Schroder, "ROSIS (Reflective Optics System Imaging Spectrometer) - A candidate instrument for polar platform missions, in Proc. SPIE 0868 Optoelectronic technologies for remote sensing from space, J. Seeley and S. Bowyer, Eds., pp 8, 1988.

[10] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in 31st International Conference on Neural Information Processing Systems, Long Beach, California, 2017