Data Driven Joint Hyperspectral Band Selection and Image Classification

Robiulhossain Mdrafi and Ali Cafer Gurbuz
Department of Electrical and Computer Engineering
Mississippi State University, Starkville, MS 39762
Email: rm2232@msstate.edu and gurbuz@ece.msstate.edu

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883821

***Abstract*—Hyperspectral sensors acquire data with a large number of spectral bands. These large number of bands make the processing computationally expensive and difficult in many real-world applications. In addition, with the spatial dimensions, the volume of the data creates problems for cases where the applications permit only limited resources both in terms of hardware computational and storage requirements. To avoid these limitations, band selection plays very pivotal role for many applications. Existing techniques utilize redundancy, clustering, sparsity, ranking type criteria for band selection. We propose an end-to-end deep learning pipeline together with a constrained measurement learning structure to select bands in a data driven manner to optimize directly the final task, which is the classification accuracy for this paper. Our results on a publicly available hyperspectral dataset show that the proposed data-driven approach provides higher classification accuracy compared to the existing state-of-art methods for the same number of bands utilized.**

***Index Terms*—Hyperspectral, Band Selection, Deep Learning, Convolutional Neural Network, Gumbel-Softmax, measurement learning.**

I. Introduction

Hyperspectral sensors acquire the spectral reflectance of the ground objects using hundreds of bands. While this provides important spectral information on a given scene that can differentiate objects, the number of bands together with spatial dimensions produces a huge volume of data, which is the hyperspectral image (HSI). This large stream of data creates high computational cost, induces 'curse of dimensionality' for a low number of training samples in high dimensional HSI data with redundant correlated information among the bands. Hence, the classification of objects in a given HSI scene degrades and may not be implementable on the fly in various real-world applications. To overcome these limitations, band selection (BS) aims to select a smaller subset of bands that captures the most relevant information.

Over the years, various BS methods have been presented in the literature. A review of these methods can be found in [1]. BS approaches are categorized as -ranking based, searching-based, clustering-based, sparsity-based, embedded learning-based, and hybrid schemes. Ranking based BS techniques select bands based on the ranking of the bands sorted by the selection score where the selection score prioritizes the selection of the top bands. Among them Maximum variance PCA (MVPCA) [2] ranks and selects the bands based on the variance of principal components obtained from the eigen value decomposition of the covariance matrix of the given HSI data. Fast density peak-based clustering (FDPC) [3] method finds the cluster centers as the distance between all pairwise bands to find the independent density peaks that corresponds to the selected bands. The main advantage of these methods are that they are computationally faster, but they fail in terms of giving enhanced classification where more sophisticated patterns are observed in the HSI data.

To select the subset of bands, searching based selection approaches aims to find the suitable and best spectral information by satisfying an optimization criterion. Among them linear prediction (LP) [4] selects most unique bands based on the similarity between a single band and multiple bands. The clustering based approaches split the bands into a set of clusters where the bands from each cluster are selected from the similarity measures [5]–[7]. Sparsity based methods like improved Sparse Subspace Clustering (ISSC) [8] finds the bands based on the notion that smaller dimensional components i.e. subspaces similarity can help us to distinguish the set of bands. In embedded learning, BS is incorporated into the optimization of the specific application models. For example in recursive feature elimination-SVM, weights calculated in SVM training are later used as ranking criteria to remove redundant bands [9]. In hybrid-based BS, combination of previously stated methods are utilized to find the most suitable form of bands [10].Although existing BS techniques provide enhanced performance, they are not directly optimizing the task metric such as classification accuracy, and they are not data-driven and automatic. In general, similarity or redundancy based information is utilized to preserve the significant spectral information which could only be a sub-optimal indicator of final task performance.

To this end, we propose an end-to-end data driven pipeline for joint band selection and classification where the architecture minimizes the classification loss and at the same time learn a constrained measurement mask to select the optimal bands. We utilize a probability mask to generate an initial estimate of the bands to get a proxy HSI data with only selected bands that is fed into a deep neural network based classification part to get the final classification score of the

This work was supported by the National Science Foundation under CAREER Grant No.2047771.

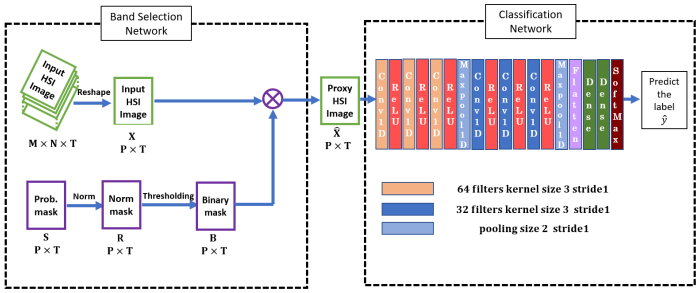

Fig. 1: Block diagram of the proposed method for T total number of bands.

input HSI data. From the classification loss in the pipeline, with the help of back-propagation, we learn a binary mask that satisfies a constraint for a given number of bands and generates the optimal bands for the problem at our hand. Our initial results shows better classification accuracy than state-of-the-art band selection based HSI classification methods.

The rest of the paper is organized as follows. Section II provides the theory and the implementation details of the proposed model for learning based joint BS and classification. Section III demonstrates the experimental settings and results. A concluding remark about the work and its extension is drawn in Section IV.

II. Proposed Method

We propose an end-to-end network that jointly learns a constrained measurement mask to select bands and a classification architecture to achieve the final goal of object classification. The general block diagram of the proposed method is shown in Fig. 1. The proposed data driven band selection/classification architecture can be decomposed into two key parts- band selection and classification networks. As shown in the diagram, band selection part takes the input HSI data and generates a proxy estimate of HSI signal via learnable masks, which is passed into a deep neural network based classification part to get the final classification label of the given pixel of the HSI data. A more detailed descriptions of the network regarding the operation of these parts are provided in the following subsections:

A. Band Selection Network

In the band selection architecture, the HSI data $\mathbf{X}$ of dimension $M \times N \times T$ enters into pipeline as the input where M, N, denote the height, width of the spatial dimensions and T denotes the total numbers of bands respectively. Since, goal of classification task is to mainly categorize each pixel into varying object classes; we reshape the dimension of $\mathbf{X}$ as $P \times T$ where $P = M \times N$. Band selection network takes another input of probability mask $\mathbf{S}$ parameterized by $\mathbf{X}$ such that $\mathbf{S} = \sigma_t(\mathbf{X})$ where each point in $\mathbf{S}$ takes non-negative continuous values i.e. $\mathbf{S} \in [0,1]^T$. Here, we define σ_t as an element-wise sigmoid function i.e. for each pixel i of $\mathbf{X}$, $\mathbf{S}_i = 1/(1 + exp(-t\mathbf{X}_i))$ where t defines the slope of the sigmoid and acts as a hyperparameter in the pipeline. Since the value of $\mathbf{S}$ lies in the region $[0,1]^T$; hence, we can realize a Bernoulli distribution at each point of $\mathbf{S}$. If we draw binary realizations from $\mathbf{S}$, we will find the mask $\mathbf{B} \in \{0,1\}^T$ such that $\mathbf{B} \sim \prod_{i=1}^{T} \beta(\mathbf{S}_i)$ where $\beta(s)$ represents the Bernoulli random variable with parameter s. For each pixel, the obtained binary mask $\mathbf{B}$ has value 1 for the bands that are selected and 0 for the bands that will not be selected. Hence, we aim to solve the following joint optimization problem for a selected band ratio of α with a constraint $\frac{1}{T}\|\mathbf{S}\|_1 = \alpha$:

$$\{\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}\} = \arg\min_{\mathbf{S},\mathbf{\Theta}} \mathbb{E}_{B \sim \prod_{i=1}^{T} \beta(\mathbf{S}_i)} \mathcal{L}(f_{\mathbf{\Theta}}(\mathbf{B} \otimes \mathbf{X})) \quad (1)$$

Here, $f_{\mathbf{\Theta}}$ is the classification network with parameters $\mathbf{\Theta}$ where $\widehat{\mathbf{X}} = \mathbf{B} \otimes \mathbf{X}$ is the input to it. Here, $\otimes$ denotes the pointwise multiplication. $\mathcal{L}$ refers to cross-entropy loss between the predicted pixel label and the ground truth one. The selected band ratio constraint ensures that the binary mask $\mathbf{B}$ has an approximate value of α which is the value of selected bands. Here as we see from (1) that the loss function takes the expectation over the Binary mask $\mathbf{B}$; hence, via approximation of the expectation by Monte-Carlo averaging, we get:

$$\{\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}\} = \arg\min_{\mathbf{S},\mathbf{\Theta}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(f_{\mathbf{\Theta}}(\mathbf{b}^{(\mathbf{k})} \otimes \mathbf{X})) \quad (2)$$

Here, $\mathbf{b}^{(k)}$ are the independent realizations drawn from the $\prod_{i=1}^{T} \beta(\mathbf{S}_i)$ distribution. The (2) takes the same form of variational autoencoder (VAE) in [11] where the authors use the re-parameterization trick to rewrite the (2) as:

$$\{\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}\} = \arg\min_{\mathbf{S},\mathbf{\Theta}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(f_{\mathbf{\Theta}}((\mathbf{U}^{(\mathbf{k})} \leq \mathbf{S}) \otimes \mathbf{X})) \quad (3)$$

Here, $\mathbf{U}^{(k)}$ deduces the independent identical realizations from $\prod_{i=1}^{T} u(0,1)$, which is a set of random uniform variables varying from 0 to 1 i.e. $[0,1]$. Thus the inequality tells us that if the inequality constraint is satisfied, then the result of this inequality would be 1.0 else 0.0. Since this inequality only involves with the probabilistic mask $\mathbf{S}$ and the random independent realizations; hence, the thresholding operation only effects these distributions. Since the whole optimization criteria in eqn. (3) deals with the discrete representations and the thresholding operation is non-differentiable; hence, the whole pipeline will not be end-to-end in this case. To make the whole loss function for the HSI band selection and classification task differentiable, we replace the thresholding operation with another element wise sigmoid function σ_r with slope r. Thus the obective function will take the form with the constraint $\frac{1}{T}\|\mathbf{S}\|_1 = \alpha$ as:

$$\{\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}\} = \arg\min_{\mathbf{S},\mathbf{\Theta}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(f_{\mathbf{\Theta}}(\sigma_r(\mathbf{S} - \mathbf{U}^{(\mathbf{k})}) \otimes \mathbf{X})) \quad (4)$$

Using this replacement approach helps us to use the trick of Gumbel-softmax [12] and concrete distributions [13] to train the proposed pipeline. Another issue is to maintain the

constraint $\frac{1}{T}\|\mathbf{S}\|_1 = \alpha$. To achieve this, we use a normalization (norm) layer that helps to rescale the value of $\mathbf{S}$. This rescaling satisfies the constraint. The norm. layer is defined as:

$$Norm_\alpha(\mathbf{S}) = \begin{cases} \frac{\alpha}{\overline{s}}\mathbf{S} & \text{if } \overline{s} \geq \alpha \\ 1 - \frac{1-\alpha}{1-\overline{s}}(\mathbf{1} - \mathbf{S}) & \text{otherwise} \end{cases} \tag{5}$$

Here, $\overline{s}$ is the average of pre-normalization of probabilistic mask S i.e. $\overline{s} = \frac{\|\mathbf{S}\|_1}{T}$. It can also be seen that eqn. (5) gives us $Norm_\alpha(\mathbf{S}) \in [0,1]^T$ and $\|Norm_\alpha(\mathbf{S})\|_1/T = \alpha$. Using this normalization layer, we can write our final objective function as:

$$\{\widehat{\mathbf{S}}, \widehat{\mathbf{\Theta}}\} = \arg\min_{\mathbf{S},\mathbf{\Theta}} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}(f_{\mathbf{\Theta}}(\sigma_r(Norm_\alpha(\mathbf{S})) - \mathbf{U^{(k)}}) \otimes \mathbf{X})) \tag{6}$$

It can be seen that, in accordance with the block diagram of Fig. 1 and the band selection network, we obtain probability mask $\mathbf{S}$ by using element wise sigmoid with slope s. Then, we get the rescaled mask $\mathbf{R}$ from the norm layer $Norm_\alpha$ i.e. $\mathbf{R} = Norm_\alpha(\mathbf{S})$. Using thresholding operation using another element wise sigmoid with slope r results the binary mask $\mathbf{B} = \sigma_r(Norm_\alpha(\mathbf{S})) - \mathbf{U^{(k)}}$. Once we get $\mathbf{B}$, we get the proxy band selected data $\widehat{\mathbf{X}}$ as the element wise multiplication of the binary mask $\mathbf{B}$ and $\mathbf{X}$. Although the dimension of $\widehat{X}$ is exactly same as of $\mathbf{X}$, only columns corresponding to 1.0 in the binary mask $\mathbf{B}$ are used as the selected bands from the total bands. Hence, the rest of columns the will be zero as they are also zero in the obtained binary mask $\mathbf{B}$. Once, we receive the band selected proxy HSI data $\widehat{X}$, it is then fed into the classification network as shown in the Fig. 1.

B. Classification Network

As shown in the Fig. 1, the classification part takes the band selected proxy HSI data $\widehat{X}$ and gives us the prediction of the given pixel label $\widehat{y}$ in the HSI data. Since we are performing pixel-wise classification; hence, we are interested only on extracting features spatially from the given HSI data. Therefore, we opt to use $1-D$ convolution for extracting features hierarchically from the BS data. We use a set of $1-D$ convolutional filters (Conv1D) with ReLUs. First set of convolutional layers consists of three Conv1D layers where each outputs 64 filters with a stride of 1 and kernel length of 3 followed by ReLU activation. With same stride, kernel length and activation, each Conv1D layer in the second set of Conv1D layers outputs 32 filters. Next, a max-pooling layer (Maxpool1D) is used to downsample the size of the extracted features. This downsampling apporach helps us to get a hierarchical representation of the features. The output of the second set of Conv1D layers are flattened before feeding them into a set of fully connected dense layers. First dense layer consists of 25 output neurons with ReLU activation while second one outputs the number of classes in the given HSI data. The output of second layers passes through soft-max function to give us the probable class distribution for the pixel of given HSI data. Once we find the soft-max of the given sample, we can write the classification loss function $\mathcal{L}$ is the cross entropy loss that is defined as

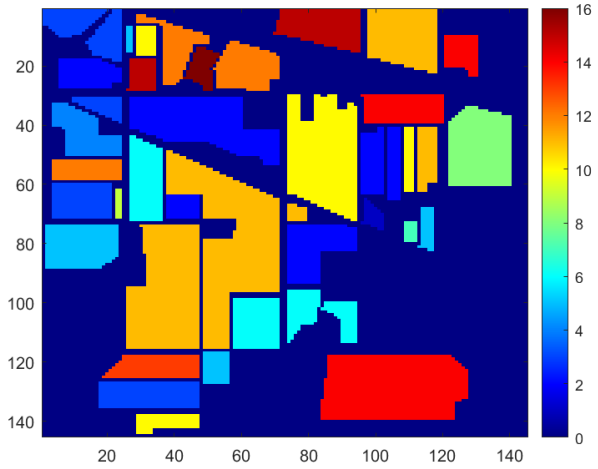

Fig. 2: Ground truth of the Indiana Pines dataset [Color bar showing different classes in the scene]

$$\mathcal{L}(\widehat{l_i}, l_i) = -\sum_{i=1}^{P}\sum_{g=1}^{C} l_{i,g} log Soft(\widehat{l_{i,g}}). \tag{7}$$

where $Soft(\widehat{l_{i,c}})$ is the soft-max layer output that gives the probability that sample i belongs to class g. Here, $\widehat{l_i} and l_i$ represents predicted and ground truth respectively. Once we have the predict the probabilities of all class labels, class label is declared as the one that corresponds to the maximum value i.e. $\widehat{y_i} = \arg\max_i \widehat{l_i}$. Since in the whole network where BS and classification stages are learned with minimizing the classification loss, the selected bands are learned to optimize classification performance and hence, we opt to name our network as Measurement learning based Band Selection (MLBS).

III. Experimental Settings and Results Analysis

In this work, we use a publicly available Indian Pines dataset [14] provided by Purdue University. This dataset was collected in a test site in Indiana via AVIRIS airborne sensor. Initially, the collected HSI data had the shape of $145 \times 145 \times 224$ where 224 denotes the total number of bands. The band number was further reduced to 200 by eliminating corrupted bands due to water absorption and radiometric corrections. The HSI data has 16 classes of ground objects in the imaging scene. It contains a total of 10249 samples with classes being alfalfa, no-till corn 1, minimal-till corn, corn, grass/pasture, grass/trees, mowed grass/pasture, windowed hay, oats, no-till soybeans, minimal-till soybeans , clean soybeans, wheat, woods, building/grass/tree drives and stone/steel towers. The ground truth of the Indian pines dataset is given in Fig. 2. We use 10% of the data from all classes for training, and the rest for the testing. We ran the proposed network for 10 times independently to get the average performance measures.

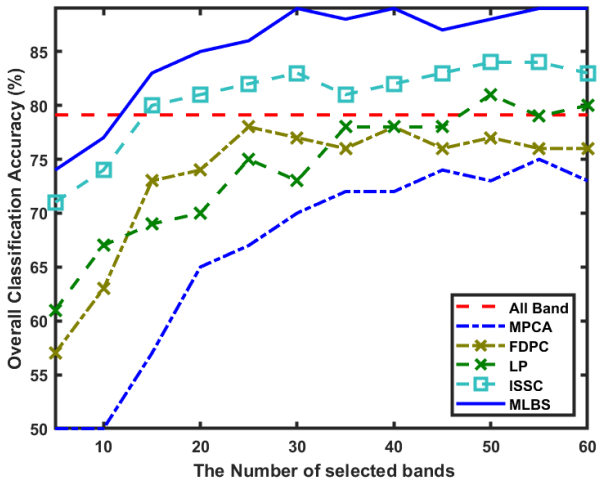

Fig. 3: Comparison of overall classification accuracy for compared methods as a function of number of selected bands.

A. Implementation of the proposed network

For implementing the probability mask $\mathbf{S}$, we use the value of slope $t = 5$ for σ_t to squash the value of $\mathbf{X}$ to the range $[0, 1]$. For approximating the thresholding operation by σ_r, we use the value of slope $r = 200$. These values are selected based on the grid-search technique. We use gradient descent based on adaptive moment estimation (ADAM) optimizer with a batch size of 16 for a varying learning rate from 0.1 to 0.0001 to determine the network parameters via Keras API.

B. Results Analysis

In this work, we use overall accuracy (OCA), average accuracy (ACA), and kappa statistics (KC) as the measures to evaluate the performance of the proposed method. For comparison purposes, we report the results of state-of-the-art band selection techniques MVPCA [2], FDPC [3], LP [4], and IISC [8] under same number of selected bands.

We show overall classification accuracy as a function of number of selected bands for compared approaches in Fig. 3. We can see that the proposed MLBS approach outperforms compared state-of-the-art approaches providing higher overall classification accuracy for all tested number of selected bands. We can also see that around $30(\alpha = 0.15$ in our case) selected bands, performance of most methods become flatter. For this number of selected bands, the proposed MLBS results around 8% more accuracy than the closest compared approach. Here, the results of all methods compared for 30 selected bands are provided in the Table I with respect to OCA, ACA, and KC for all classes in the dataset. The results reported in the table also show that the proposed MLBS approach results in superior classification performance.

TABLE I: Classification results of different band selection methods for 30 selected bands.

Class Name	Method					
	MVPCA	FDPC	LP	ISSC	MLBS	ALL
ACA	65.99	68.09	67.04	76.85	**82.88**	72.65
OCA	70.18	77.45	73.5	81.61	**89.08**	79.12
KC	65.83	74.22	69.56	78.98	**81.14**	76.05

IV. Conclusions and Future Work

In this work, a data driven deep neural network based pipeline is proposed to jointly select the bands from the hyperspectral data in order to minimize classification loss. This way different from existing approaches who look for similarity or redundancy related metrics to select bands, the bands that directly optimize the final task (i.e., classification) related cost are selected. It has been shown on a publicly available dataset that the proposed band selection method outperforms state-of-the-art approaches for hyperspectral object classification. The future work will provide enhanced analysis and comparisons of the proposed approach under varying scenarios .

References

[1] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 118–139, 2019.

[2] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE transactions on geoscience and remote sensing*, vol. 37, no. 6, pp. 2631–2641, 1999.

[3] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 88–102, 2015.

[4] H. Du, H. Qi, X. Wang, R. Ramanath, and W. E. Snyder, "Band selection using independent component analysis for hyperspectral image processing," in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, pp. 93–98, IEEE, 2003.

[5] M. Ahmad, D. I. U. Haq, Q. Mushtaq, and M. Sohaib, "A new statistical approach for band clustering and band selection using k-means clustering," *Int. J. Eng. Technol*, vol. 3, no. 6, pp. 606–614, 2011.

[6] T. Imbiriba, J. C. M. Bermudez, C. Richard, and J.-Y. Tourneret, "Band selection in rkhs for fast nonlinear unmixing of hyperspectral images," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1651–1655, IEEE, 2015.

[7] S. Li, J. Qiu, X. Yang, H. Liu, D. Wan, and Y. Zhu, "A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search," *Engineering Applications of Artificial Intelligence*, vol. 27, pp. 241–250, 2014.

[8] W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2784–2797, 2015.

[9] R. Zhang and J. Ma, "Feature selection for hyperspectral data based on recursive support vector machines," *International Journal of Remote Sensing*, vol. 30, no. 14, pp. 3669–3677, 2009.

[10] S. Li, H. Wu, D. Wan, and J. Zhu, "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowledge-Based Systems*, vol. 24, no. 1, pp. 40–48, 2011.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[13] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[14] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3," *Purdue University Research Repository*, vol. 10, p. R7RX991C, 2015.