Surrogate Contrastive Network for Supervised Band Selection in Multispectral Computer Vision Tasks

Edgar A. Bernal
Rochester Data Science Consortium
University of Rochester
260 E. Main St., Rochester, NY 14604
edgar.bernal@rochester.edu

Abstract

Computer vision techniques that operate on hyper- and multispectral imagery benefit from the additional amount of spectral information relative to those that exploit traditional RGB or monochromatic visual data. However, the increased volume of data to be processed brings about additional memory, storage and computational requirements. In order to address such limitations, a wide range of techniques for dimensionality reduction have been introduced by previous work. In this paper, we propose a framework for spectral band selection that is highly data- and computationally efficient. The method leverages a convolutional siamese network learned by optimizing a contrastive loss, and performs band selection based on the low-dimensional data embeddings produced by the network. We empirically demonstrate the efficacy of the method on an object detection task from aerial multispectral imagery. The results show that, in spite of the method's frugality, it produces very competitive band selection results against the evaluated competing techniques.

1. Introduction

Hyper- and multispectral imaging techniques aim at collecting information across the electromagnetic spectrum, thus enabling the reconstruction of the spectral signature of the scene –and objects therein– being imaged. Since real-world materials possess distinct spectral characteristics (i.e., different absorption and reflection responses across wavelengths of electromagnetic radiation), analysis of hyper- and multispectral imagery has a wide range of applications that rely on object identification and localization, including healthcare, astronomy, remote sensing, agriculture, surveillance, and the like. While acquisition and processing of a large number of bands may lead to improved performance, the increased amount of associated data can impose unwieldy storage and computational constraints. Consequently, dimensionality reduction techniques, both during acquisition/reconstruction [22, 15] and processing and analysis [24, 4, 2] have been proposed.

In this paper, we investigate methods for supervised band or feature selection, which aim at finding an optimal subset of bands by leveraging task-relevant labels associated with the imagery. More specifically, we focus on deep learning frameworks for supervised band selection. Some recent work has proposed the integration of a supervised feature selection mechanism with the decision-making algorithm [16, 17], i.e., the deep classifier. The main shortcoming of such approaches is that deep networks trained in a supervised manner tend to be data hungry, and since the band downselection is built into the network, it can only be enabled by large amounts of data and the accompanying labels. Additionally, band selection is only available after the classifier has been trained on the full set of data. We believe there is an advantage to having a band selection entity separate from the main learning framework, particularly if it can operate on reduced data corpora; this is especially true in cases where obtaining large amounts of labeled data at *full* spectral resolution is expensive or simply unfeasible. Another recent body of work [8] implements separate band selection mechanisms from classification frameworks which partially alleviates the initial data requirements; remarkably, the dimensionality reduction is carried out with the aid of deep networks, while the classification relies on shallow models. Additionally, although the target decision-making task is image classification, all spatial information is discarded when performing band selection, which is done purely based on pixel-level spectral information. We hypothesize there is an advantage to performing the band selection and the machine learning tasks on comparable domains. Further, we believe that it is more sensible to impose capacity constraints on the dimensionality reduction front-end than on the intelligence back end.

We propose a framework for supervised band selection that leverages a surrogate, low-capacity deep network employing a contrastive loss function. While the proposed band selection mechanism is classifier-agnostic, we demonstrate its performance towards the goal of object detection from multispectral imagery, arguably a more difficult one than the more commonly implemented pixel- or image-level classification decision-making tasks. The contributions of this paper can be summarized as follows:

- We introduce the use of a highly data- and computationally efficient surrogate network to aid the band selection process; the nature of the network allows it to operate on significantly reduced corpora, thus alleviating data, label and computational requirements.
- The computational efficiency of the proposed approach is accentuated by the need to train a single surrogate network, in contrast with some existing methods which require learning separate classifiers for each band selection scenario and/or each class under consideration.
- While the proposed band selection framework is classifier-agnostic, the features on which it relies are well-matched with those used by deep convolutional networks.
- We empirically demonstrate the effectiveness of the proposed framework on a building detection task carried out on the Paris portion of the SpaceNet [1] dataset.

This paper is organized as follows: Sec. 2 provides an overview of relevant literature. Sec. 3 describes the proposed discriminability metric and the network architecture leveraged in its computation. Sec. 4 describes the experimental framework and results, and Sec. 5 concludes the paper.

2. Related Work

Generally speaking, dimensionality reduction techniques for hyper and multispectral imagery fall in one of two categories [24]: feature extraction [4] and band selection (sometimes referred to as feature selection) [2]. Feature extraction approaches aim at estimating lower-dimensional representations of the signal [24, 23], while band selection techniques attempt to find an optimal subset of bands based on some criterion [2, 16]. The main advantage of band selection techniques is that the resulting compressed representation maintains the physical interpretation present at the source [14].

Band selection approaches can in turn be classified as unsupervised [7, 13, 20] and supervised [8, 3, 17], while some frameworks combine both supervision modalities [21, 5]. As their name implies, unsupervised techniques do not require labels to operate; due to their task-agnostic nature, they can be applied in a wide variety of scenarios. However, their performance suffers relative to supervised approaches, which take advantage of task-relevant labels. When the band selection is effected simultaneously with the main analysis task, the band selection framework falls in the wrapper category [19]; in contrast, when the band selection is performed before the main task, the approach is termed a filter. While filter approaches tend to suffer in terms of performance, their advantage is that they can be trained independently of the main classifier, and thus, are not subject to the same computational and data volume requirements. This is particularly beneficial in cases where the main classifier is a supervised deep framework, as those tend to be data- and computationally demanding. In this paper, we propose a filter approach which affords a significant savings in data and computational requirements relative to the deep network used for the object detection task.

Recently, in line with the increasing popularity of deep frameworks that can be trained in an end-to-end fashion, supervised band selection approaches that are built into the network architecture have been proposed [16, 17]. The authors of [16] implement an attention mechanism on top of the classifier which identifies the most informative bands in a dataset during the training process for a given task. In contrast, the framework from [17] identifies the most influencial bands by measuring the contribution of each band to the overall loss being optimized. Unfortunately, both of these methods are wrappers, and as such require the availability of a large, labeled set of imagery, since they tap into the iterative training process of the network used for the main task. This partially defeats the purpose of a dimensionality reduction scheme whose aim is to reduce data requirements and computational loads throughout the various stages of the process. With this in mind, we propose a filter-type, classifier-agnostic band selection approach that is able to operate on significantly reduced corpus of labeled data.

The method from [8] is most similar to the one proposed herein, as it operates independently of a classifier and leverages a neural network. However, the identified features are not necessarily well matched to the implemented classifiers, which are shallow in nature (i.e., a k-nearest neighbors clustering algorithm and a regression and a classification tree). While the architecture of our proposed band selection network is different than that of the classifier used to demonstrate its efficacy –a natural consequence of its decreased capacity and lower data requirements–, the band selection is performed on features that more closely match the domain of operation of the classifier.

In order to enable data-efficient band selection, we leverage the siamese network architecture [12], first introduced to be employed in scenarios where data scarcity is an issue, including few-shot learning tasks. We extract band-dependent representations of input samples and measure the

discriminability of each band from the normalized inter-class separation of its resulting embeddings. We train the network with a contrastive loss [9], which encourages samples from the same/different class to be mapped onto nearby/distant points in the embedding manifold. We demonstrate the performance of the proposed framework on the task of object detection. To that end, we leverage the publicly available SpaceNet [1] dataset, and modify the architecture of the YOLO network [18] to support processing of multispectral imagery.

3. Estimating Band Discriminability

In this section, we introduce the framework used to estimate the discriminative power of each of the bands available for the ultimate object detection task. As stated, the desired characteristics of the band selection architecture are data and computational efficiency, as well as the ability to operate on a feature space that resembles that of the classifier, in our case, the fully convolutional YOLOv2 network. Given these considerations, we leverage the data efficiency of the siamese architecture [12] and implement a convolutional version of it. Fig. 1 illustrates the architecture of the network. Each branch in the model consists of a sequence of convolutional layers followed by PReLU [11] activations and max-pooling layers. The two-dimensional feature map output by the convolutional stages is flattened into a vector that goes through two additional fully connected layers, resulting in a vectorial embedding of length 256.

As illustrated in Fig. 1, the value of a contrastive loss function [9] is computed between the siamese activations at the last layer. Specifically, let N denote the minibatch size, $(x_1^{(i)}, x_2^{(i)})$ the length-m network activations (in our case, $m = 256$) for the i-th sample pair in the batch, and $y^{(i)}$ their corresponding label: $y^{(i)} = 1$ if both samples belong to the same class, and $y^{(i)} = 0$ otherwise. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the $N \times m$ matrices obtained by stacking the set of N activations, $x_1^{(i)}$ and $x_2^{(i)}$ respectively, and $\mathbf{Y}$ be the $N \times 1$ vector containing the correspondingly stacked labels, $y^{(i)}$, for $i = 1, 2, \ldots, N$. The contrastive loss takes the form:

$$\mathcal{L}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} \Big[(1 - y^{(i)}) \mathcal{D}(x_1^{(i)}, x_2^{(i)})^2 + y^{(i)} \max(0, \mathcal{M} - \mathcal{D}(x_1^{(i)}, x_2^{(i)})) \Big] \tag{1}$$

where $\mathcal{D}$ denotes a distance metric, in our case the Euclidean distance, and $\mathcal{M}$ a scalar margin selected via hyperparameter optimization.

We extract 200 8-channel patches, each with 64×64 pixels, chosen randomly from the training set. We separate the 8-channel patches into 8 single-channel images and feed them to the siamese network from Fig. 1. The labels associated with the images reflect whether they were extracted from the 'background' class or from the 'object' class. Instead of implementing a generic contrastive learning procedure wherein random pairs of patches are fed to the network, we constrain the pairs of patches to belong to the same band. In other words, for any particular training pair, the network sees a pair of image patches extracted from the same spectral channel. This is because we want to be able to measure feature discriminability at the band-level in order to perform band selection; with this goal in mind, learning attributes that differentiate images from different spectral bands is not relevant.

Once the network is trained, we estimate the discriminative abilities of each band by measuring the degree of separation between the learned embeddings of the training images corresponding to each class for each spectral band individually. Specifically, we measure between- or inter-class variance for features from band j, denoted as $\sigma_b^2(j)$ according to:

$$\sigma_b^2(j) = \omega_0(j)\omega_1(j)(\mu_0(j) - \mu_1(j))^2 \tag{2}$$

where $\omega_i(j)$ and $\mu_i(j)$, for $i = 0, 1$, denote the fraction of samples and the mean embedding vector for class i, respectively, for the training data corresponding to band j. $\sigma_b^2(j)$ measures how far apart the clusters corresponding to the embeddings of the different classes are located with respect to each other; however, it doesn't consider the spread of each individual cluster. Discriminative features should push the embeddings of samples from different classes apart, while pulling the embeddings of samples from a given class close together. The spread of the clusters of embeddings of samples from individual classes for features from band j can be measured via the within- or intra-class variance, denoted as $\sigma_w^2(j)$ and computed according to:

$$\sigma_w^2(j) = \omega_0(j)\sigma_0^2(j) + \omega_1(j)\sigma_1^2(j) \tag{3}$$

where $\sigma_i^2(j)$, for $i = 0, 1$, denotes the variance of the embedding vectors for class i and band j. We measure the discriminative capabilities of band j by computing the ratio of the inter-class variance to the total variance, according to:

$$\Delta(j) = \left\| \sigma_b^2(j) \oslash (\sigma_b^2(j) + \sigma_w^2(j)) \right\| \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\oslash$ denotes the element-wise, or Hadamard division between vectors.

Eq. 4 enables estimation of the discriminative ability of features extracted from each individual band j. In order to select a subset of bands, we need to take into account the interaction between the bands. It is well known that good features are not only discriminative but also largely

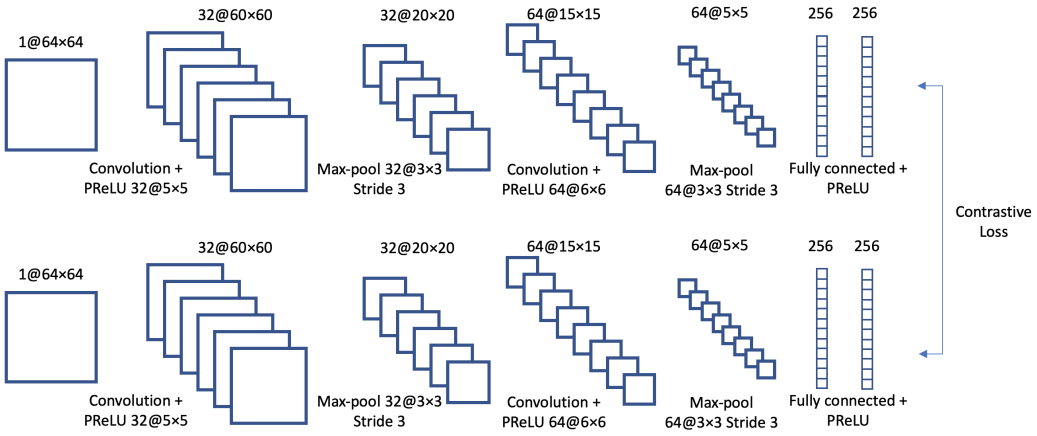

Figure 1: Siamese network architecture.

uncorrelated with each other [10]. With this in mind, cross-correlation coefficients across the embeddings corresponding to different bands are computed. The band selection process consists in ranking the bands according to their discriminative abilities (based on Eq. 4) and selecting bands whose cross-correlation coefficient with bands already selected does not exceed a predetermined threshold. If the full set of bands is processed without having selected the number of required bands, the remaining bands are selected from the subset of bands with the largest discriminability. Alternatively, the cross-correlation coefficient threshold can be iteratively increased.

The described band selection method assumes that the classification task involves two classes. In our case, the classes of interest are 'building' and 'background'. In a C-class classification task, where $C > 2$, the siamese network is trained in a similar fashion. The resulting embeddings are analyzed in a slightly different manner: in a class-level selection framework, $C - 1$ one-vs-all discriminative metrics for each band (Eq. 4) and $C - 1$ cross-correlation matrices are computed. The band selection is then performed independently for each class. In a global selection framework, the selection process can operate on an aggregate per-band discriminative metric and a single cross-correlation matrix. Note that some supervised band selection methods in existing literature require the construction of one framework per band and/or one framework per class, which can quickly become unmanageable.

4. Experiments

4.1. Dataset

The effectiveness of the proposed band selection method was tested on an object detection task on multispectral imagery from the Paris portion of the SpaceNet [1] dataset. The SpaceNet Paris corpus includes approximately 1,030km^2 full-resolution 30cm imagery collected from DigitalGlobe's WorldView-2 commercial satellite and includes 8-band multispectral data. The dataset also includes 23,816 building labels extracted from the imagery. The 8-band, multispectral images include the following bands: Coastal Blue, Blue, Green, Yellow, Red, Red Edge, Near Infrared 1 (NIR1), and Near Infrared 2 (NIR2). The object detection performance was tested on a total of 4,591 8-channel images, each with 416×416 pixels per channel. An 80/20 partition scheme was implemented, meaning that the training set comprised 3,674 images while the test set included 917. Fig. 2(a) shows a sample image and Fig. 2(b) shows its corresponding annotation file; these images illustrate the type of data used in the object detection task. As stated, the band selection was effected on a subset of 200 8-channel patches, each with 64×64 pixels, representing a mere 0.1% of the total amount of training data available for the detection task. Fig. 2(c) shows two sample patches used for band selection purposes, one for the building class (green box) and one for the background class (red box).

4.2. Experimental Setup

A modified YOLO [18] network was used as the object detector of choice. Support of multi-channel imagery was achieved by incorporating the Geospatial Data Abstraction Library (GDAL) [6] which enables reading and writing raster and vector geospatial data formats. The detection framework was executed on a machine with an Intel Core i7 quad-core processor, 64GBytes of RAM, 3TBytes of storage and dual nVidia GeForce GTX 1080 GPUs, each with 11GBytes of video RAM. The machine ran Ubuntu 18.04, with nVidia driver version 415.27 and CUDA version 10.0. The band selection framework, in contrast, was executed on a MacBook Pro with an Intel Core i7 quad-core proces-

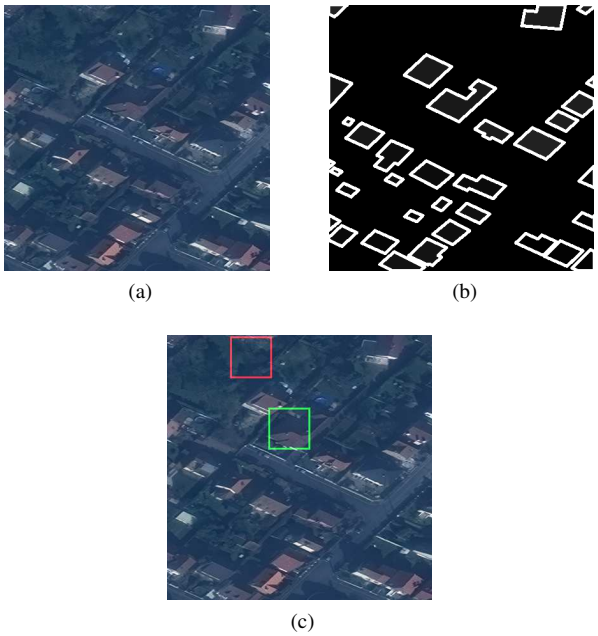

Figure 2: Example images from SpaceNet: (a) RGB image and (b) corresponding annotation image used for object detection purposes; and (c) sample positive and negative extracted patches used for band selection purposes.

sor and 32GBytes of RAM and 500GBytes of storage. In line with our view that band selection should be a resource-constrained (both computationally and in terms of data) process, no GPU acceleration was leveraged for band-selection purposes.

4.3. Performance Metrics

We evaluate the performance of a given detection algorithm based on the mean average precision (mAP) metric, which has become a standard in object detection challenges. The mAP is the average of the maximum precisions across different recall values, obtained as the detection threshold is varied. Since it is derived from the Precision-Recall curve, we sometimes include the curve to facilitate visual examination of performance, particularly when multiple systems are being compared. Precision and Recall are functions of the number of true positives, false positives and false negatives. We determine a detection to be a true positive if its intersection over union (IoU) with the ground truth annotation is larger than 0.5 and a false positive otherwise. A false negative is an annotation without a corresponding detection with larger than 0.5 IoU. Fig. 3 illustrates the dependence of the triggered detections on the value of the detection threshold as a building detection algorithm is applied to the image from Fig. 2(a). As expected, the number of detections decrases as the detection threshold is incrased, which results in reduced true and false positives, as well as increased true and false negatives.

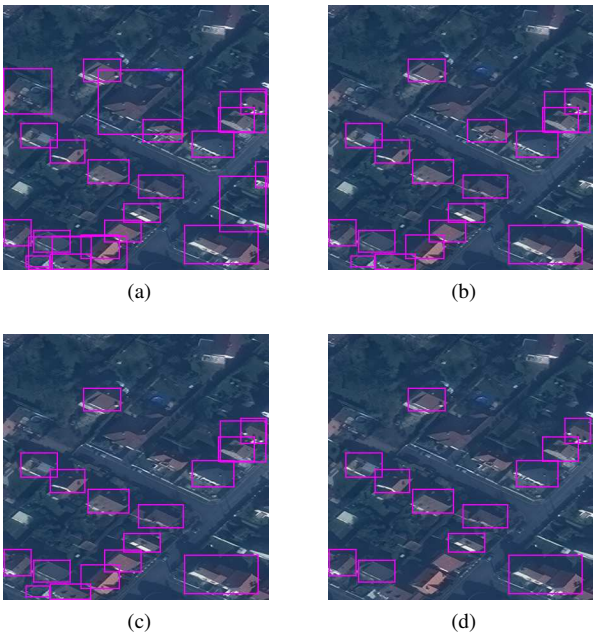

Figure 3: Triggered detections as threshold is increased.

4.4. Preliminary Results

By way of an initial sanity check, we computed the performance of an eight-band detection system as the learning process took place. The curves in Fig. 4 show the Precision-Recall behavior of the system after having learned from 4, 6, 8 and 10 thousand labeled samples. It can be seen that, as expected, the area under the curve (or, equivalently, the mAP) increases as the learning process progresses.

Next, we compared the performance of systems naïvely operating on one-, three-, six- and eight-band imagery. The selection criterion in this case was approximately uniform sampling of the spectrum. The bands selected were Band 4 (Yellow) for the one-band case, Bands 2, 4 and 6 (Blue, Yellow and Red Edge) for the three-band case and Bands 1, 2, 4, 5, 7 and 8 (Coastal Blue, Blue, Yellow, Red, NIR1 and NIR2) for the six-band case. No principled band selection approach was carried out for this experiment. Fig. 5 contains the results. As expected, having access to larger number of bands affords a system improved detection performance.

4.5. Band Selection Results

In order to test the efficacy of the proposed band selection framework, YOLO-based object detectors were trained on modified datasets containing one-, three- and six-channel

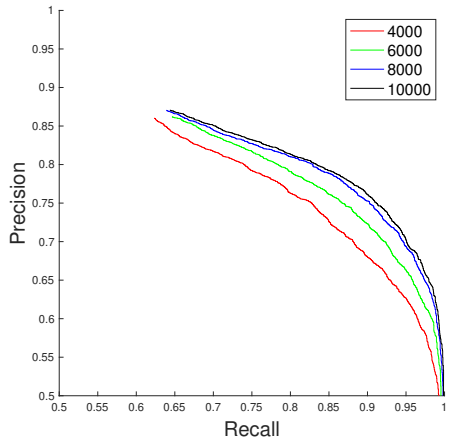

Figure 4: Precision-Recall curves for an eight-band detection system as a function of iteration number.

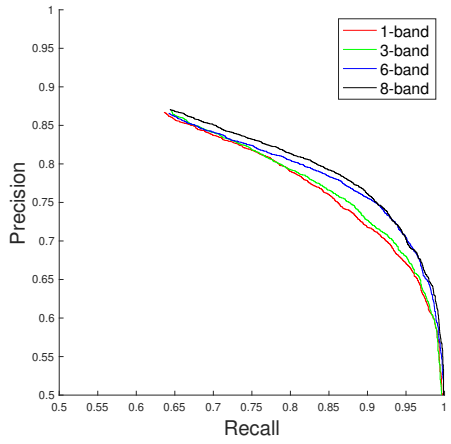

Figure 5: Precision-Recall curves for one-, three-, six- and eight-band detection systems.

imagery. The choice of channels in the imagery were provided by three different methods, namely: (i) the naïve selection method described in Sec. 4.4 where bands were selected to achieve approximate uniform spectrum sampling; (ii) the method from [8], which we consider to be state-of-the-art and closest in nature to our proposed method; and (iii) our proposed method.

4.5.1 Discriminability and Cross-correlation Estimates

The algorithm described in Sec. 3 on a subset of 1600 one-channel patches of dimensions 64×64 pixels (resulting from spectrally splitting 200 8-channel patches of the same spatial dimensions) extracted from training set imagery, 800 corresponding to each class, namely 'building' and 'background'. The curve in Fig. 6 illustrates the estimated discriminative power for each of the eight bands under consideration as estimated by Eq. 4. Band 7 (NIR1) and Band 1 (Coastal Blue) were deemed to be the most and least discriminative bands, respectively, by the proposed surrogate network approach. The heat map in Fig. 7 shows a pseudo-colored version of the cross-correlation matrix $\mathbf{R} \in \mathbb{R}^{8\times 8}$ for the feature embeddings across all bands: entry $\mathbf{R}(i,j)$ corresponds to the cross-correlation coefficient between bands i and j, for $i, j = 1, 2, \ldots 8$; note that $\mathbf{R}(i,j) == \mathbf{R}(j,i)$. There is a strong tendency for neighboring bands to be correlated with each other, as intuition would indicate.

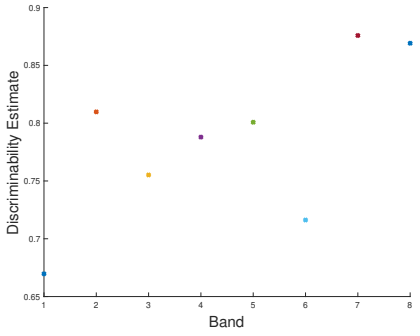

Figure 6: Estimated discriminative power of each of the spectral bands in the Paris dataset (larger is better).

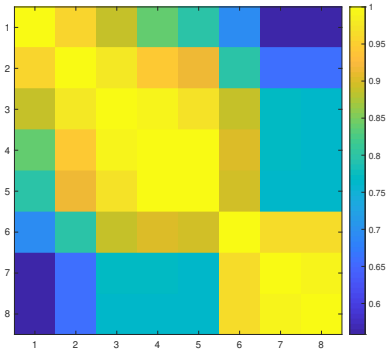

Figure 7: Cross-correlation matrix for the feature embeddings across all bands (larger means higher correlation).

4.5.2 Single-Band Results

Our proposed method selected Band 7 (NIR1), as evidenced by the results illustrated in Fig. 6. The naïve method se-

Table 1: Detection performance in mAP (larger is better)

T/M	One Band	Three Bands	Six Bands
Uniform	0.8288	0.8257	0.8385
Ref. [8]	0.8263	0.8278	0.8371
Proposed	**0.8359**	**0.8372**	**0.8395**

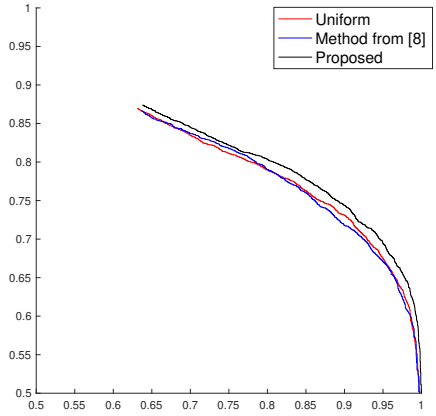

Figure 8: Precision-Recall curves for competing one-band detection systems.

lected Band 4 (Yellow), while the method from [8] selected Band 3 (Green) for a single-band system. The Precision-Recall curves for these three band choices are shown in Fig. 8. The resulting mAP figures are included in Table 1 (leftmost column); the table entries correspond to 4-run averages. It can be seen that the system based on the single band selected by the proposed framework outperforms the competing methods. Further, it can be seen that the band discriminability predictions derived from Eq. 4 and plotted in Fig. 6 correspond with the observed performance, with Band 7 (our method) outperforming Band 4 (naïve approach), and Band 4 in turn outperforming Band 3 (method from [8]).

4.5.3 Three-Band Results

Selection of three bands involves iteratively identifying the individual band with the largest discriminative ability according to Eq. 4 and with a correlation coefficient with previously selected bands smaller than a predetermined threshold. Using a threshold of 0.95, the first band selected is Band 7 (NIR1), followed by Band 2 (Blue) –Band 8 (NIR2) can't be selected given its high correlation with Band 7, in spite of its high discriminability– and Band 5 (Red). We point out that $\mathbf{R}(7,8) = 0.986$, $\mathbf{R}(2,7) = 0.660$, $\mathbf{R}(5,7) = 0.764$ and $\mathbf{R}(2,5) = 0.914$. According to

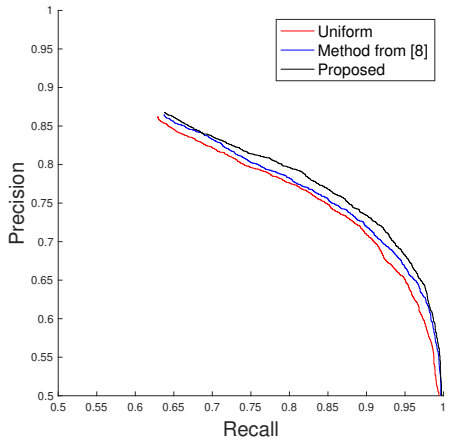

Figure 9: Precision-Recall curves for competing three-band detection systems.

the uniform sampling approach, the naïvely selected set of bands comprises Bands 2, 4 and 6 (Blue, Yellow and Red Edge). The method from [8] selected Bands 3, 5 and 8 (Green, Red, and NIR2).

The resulting mAP figures are included in Table 1 (center column) where, again, the table entries correspond to 4-run averages. The system based on the three bands identified by the proposed approach outperforms the other two sampling approaches. Fig. 9 contains the Precision-Recall curves for the three competing methods.

4.5.4 Six-Band Results

Continuing to build on the output of the three-band selection procedure, Bands 7, 2 and 5 are selected first. Since all remaining bands violate the cross-correlation constraint, Bands 8, 4 and 3 (NIR2, Yellow and Green) are selected next, in descending order of estimated discriminability. According to the uniform sampling approach, the naïvely selected set of bands comprises Bands 1, 2, 4, 5, 7 and 8 (Coastal Blue, Blue, Yellow, Red, NIR1 and NIR2). The method from [8] selected Bands 1 through 6 (Coastal Blue, Blue, Green, Yellow, Red and Red Edge). The Precision-Recall curves for these three band choices are shown in Fig. 10. The resulting mAP figures are included in Table 1 (rightmost column); as before, the table entries correspond to 4-run averages. The system based on the six bands identified by the proposed approach outperforms the other two sampling approaches, although in this case by smaller margins than in the previous scenarios. We believe this is because the band choice becomes less important as the spectral sampling scheme becomes denser and the full amount of available bands is approached.

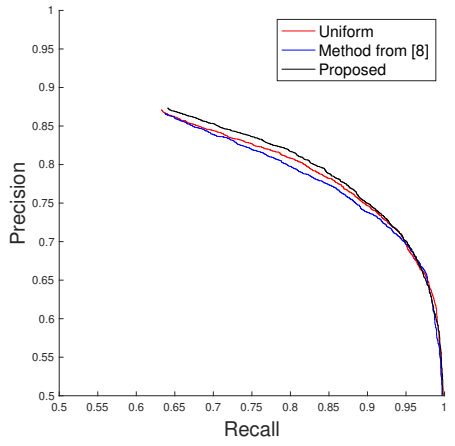

Figure 10: Precision-Recall curves for competing six-band detection systems.

4.5.5 Impact of Band Cross-Correlation

Although a rigorous study on the impact of band cross-correlation on the performance of the detection system is beyond the scope of this paper, in this section we delve further into trends observed with regards to systems where the bands in the selected set exhibit larger cross-correlation relative to competing systems. Consider the six-band choice by the method from Ref. [8] where the selected bands are contiguous, and note that its performance is worse than both of the competing methods. In contrast, for the three-band case, the same method outperformed the system operating on a set of arguably more highly correlated bands picked by the naïve approach. To further validate the hypothesis that high feature cross-correlation is detrimental to ultimate performance, we considered a hypothetical detection system based on a traditional RGB camera. We note that RGB bands exhibit high cross-correlation, being that they are close spectral neighbors, and the narrowest spectral spread of any three-band system considered. The mAP resulting from an average of 4 runs was 0.8203, lower than any other three-band system evaluated.

5. Conclusions

We proposed a highly data-efficient framework for supervised band selection in tasks related to automated analysis of hyper- and multispectral imagery. The framework comprises a surrogate convolutional siamese network optimized based on a contrastive loss function. The embeddings of the training samples learned by the surrogate network are analyzed to extract band-level discriminability and cross-correlation metrics that aid the band selection process. We demonstrated the efficacy of the proposed framework via extensive empirical evaluation on the task of object detection from aerial imagery, and compared its performance against uniform and state-of-the-art supervised sampling techniques. In spite of the ultra low data and computational requirements imposed by the approach, the results show that it can effectively identify the most relevant spectral bands given a desired number of target bands.

References

[1] Spacenet on Amazon Web Services (AWS). last modified April 30, 2018. `https://spacenetchallenge.github.io/datasets/datasetHomePage.html`. Accessed February 25, 2019.

[2] Hyperspectral band selection based on multi-objective optimization with high information and low redundancy. *Applied Soft Computing*, 70:604 – 621, 2018.

[3] X. Cao, T. Xiong, and L. Jiao. Supervised band selection using local spatial information for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters*, 13(3):329–333, March 2016.

[4] B. B. Damodaran, N. Courty, and S. Lefèvre. Sparse Hilbert Schmidt independence criterion and surrogate-kernel-based feature selection for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2385–2398, April 2017.

[5] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang. Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2956–2969, May 2015.

[6] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2018.

[7] M. Habermann. Unsupervised band selection in hyperspectral images using autoencoder. *IET Conference Proceedings*, pages 6 (6 pp.)–6 (6 pp.)(1), January 2018.

[8] M. Habermann, V. Fremont, and E. H. Shiguemori. Supervised band selection in hyperspectral images using single-layer neural networks. *International Journal of Remote Sensing*, 0(0):1–27, 2018.

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006.

[10] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, 1999.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society.

[12] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015.

[13] F. Li, P. Zhang, and H. Lu. Unsupervised band selection of hyperspectral images via multi-dictionary sparse representation. *CoRR*, abs/1802.06983, 2018.

[14] J. Li and H. Liu. Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 32(2):9–15, Mar. 2017.

[15] Q. Li and E. A. Bernal. An algorithm for parallel reconstruction of jointly sparse tensors with applications to hyperspectral imaging. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 218–225, July 2017.

[16] P. R. Lorenzo, L. Tulczyjew, M. Marcinkiewicz, and J. Nalepa. Band selection from hyperspectral images using attention-based convolutional neural networks. *CoRR*, abs/1811.02667, 2018.

[17] A. Ortiz, A. Granados, O. Fuentes, C. Kiekintveld, D. Rosario, and Z. Bell. Integrated learning and feature selection for deep neural networks in multispectral images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[18] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[19] A. H. Shahana and V. Preeja. Survey on feature subset selection for high dimensional data. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–4, March 2016.

[20] J. Tschannerl, J. Ren, J. Zabalza, and S. Marshall. Segmented autoencoders for unsupervised embedded hyperspectral band selection. In *7th European Workshop on Visual Information Processing*, November 2018.

[21] C. Wang, L. Zhang, W. Wei, and Y. Zhang. When low rank representation based hyperspectral imagery classification meets segmented stacked denoising auto-encoder based spatial-spectral feature. *Remote Sensing*, 10(2):284, Feb 2018.

[22] S. Yang, M. Wang, P. Li, L. Jin, B. Wu, and L. Jiao. Compressive hyperspectral imaging via sparse tensor and nonlinear compressed sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):5943–5957, Nov 2015.

[23] J. Zabalza, J. Ren, M. Yang, Y. Zhang, J. Wang, S. Marshall, and J. Han. Novel folded-pca for improved feature extraction and data reduction with hyperspectral imaging and sar in remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:112–122, 7 2014.

[24] W. Zhang, X. Li, and L. Zhao. Band priority index: A feature selection framework for hyperspectral imagery. *Remote Sensing*, 10(7), 2018.