SPATIAL-SPECTRAL CONTRASTIVE LEARNING FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Peiyan Guan, Edmund Y. Lam

Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam, Hong Kong SAR, China

ABSTRACT

In spite of being widely used in hyperspectral image (HSI) classification, most deep learning algorithms require plenty of labeled samples to achieve satisfactory performance. However, manual labeling is very time-consuming and laborious in practice. To solve such problem, we propose a method, called spatial-spectral contrastive learning (SSCL), to learn the representations of HSIs suitable for classification in an unsupervised manner. We demonstrate that the useful contents (e.g., semantics) are invariant under the spatial and spectral domains while the uninformative ones are usually not. Thus, we learn powerful representations that model domain-invariant information by defining a contrastive prediction task. Specifically, two signals are constructed for an HSI sample to include information of the two domains, and the representations of these two signals are then optimized to be similar, such that the domain-invariant contents are extracted. We conduct classification experiments on the learned representation with very few labels, the results of which verify the superiority of our method over the state-of-the-art techniques.

Index Terms— Unsupervised learning, contrastive learning, hyperspectral image classification.

1. INTRODUCTION

Since it is capable of capturing both spatial and spectral information of the ground objects, hyperspectral imaging technology has been broadly used in many applications, including mineral detection [1] and environmental monitoring [2]. Hyperspectral image (HSI) classification, as a key analysis method for these applications, has gained a lot of attention. Early methods propose to classify the HSIs based on hand-crafted features, which suffer from very poor performance [3]. Recently, owing to the development of deep learning in image processing [4–6], many methods have applied deep learning algorithms to solve the HSI classification and present impressive performance [7, 8]. However, the lack of labeled data still remains a main obstacle for most deep learning models. Many methods have been proposed to solve the problem, among which deep unsupervised learning is an important branch.

Deep unsupervised learning algorithms extract features from the samples without label information. One well-known tool is the autoencoder, which learns a latent encoding oriented for the input reconstruction. Numerous methods employing autoencoders for the unsupervised HSI classification have been proposed [9, 10]. Another architecture widely used is the generative adversarial network (GAN) [11, 12]. It consists of a generator and a discriminator, which compete against each other such that the generator learns the real data distribution while the discriminator learns to extract the deep features. The major drawback of these methods is that they try to code all the information of the input data. However, HSIs are in fact very low-semantic. Many contents, such as redundancies and noise, are less useful for object identification. Thus, these representations coding all the contents do not perform well on subsequent classification problems.

In contrast, we aim to recognize the contents useful for identification and learn representations that only code such contents without label information. HSI is able to capture spatial and spectral features of an object together. Although different from each other, these two types of features reflect characteristics of the same object, which demonstrate that the desired information related to semantics is domain-invariant. It also corresponds to the fact the way we sense an object does not change what it is. In contrast, the undesired contents, such as noise and redundancy, are usually independent between the two domains.

Our goal is thus to learn representations that capture the domain-invariant information. To do so, we propose a spatial-spectral contrastive learning (SSCL) framework. Specifically, we construct a pair of signals for each HSI sample to represent the two domains, respectively. Then, we learn two embedding functions such that the two signals of the same sample are projected to nearby points while those of different samples are mapped to distant points. Our main contribution is to learn powerful representations by exploring the domain-invariant information. To evaluate the quality of our representations, i.e., whether suitable for classification, we conduct classification experiments on the learned representations with limited

The work is supported in part by the Research Grants Council of Hong Kong (GRF 17200019, 17201620, 17200321) and the University of Hong Kong (104005864).

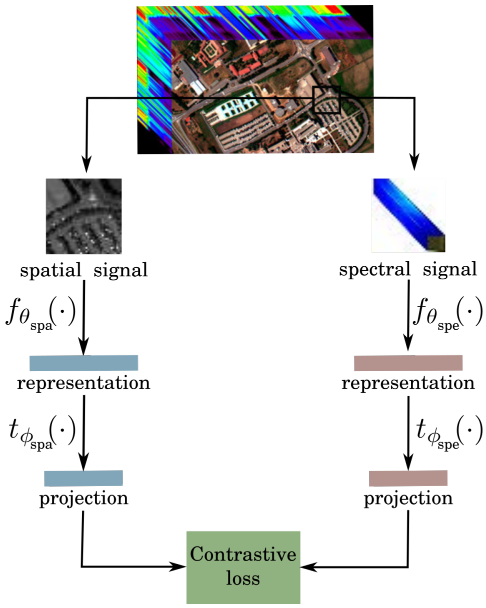

Fig. 1. Architecture of the SSCL framework.

labeled samples, the results of which illustrate the effectiveness of our method on differentiating the useful versus the undesired contents.

2. METHODOLOGY

2.1. Goal and motivation

The object of this work is to learn semantically meaningful representations of HSI samples in an unsupervised manner. Consider an HSI dataset of N unlabeled samples $X = \{x^i\}_{i=1}^N$. The sample $x^i \in \mathbb{R}^{W\times W\times C}$ denotes the ith hyperspectral cube, where C is the size of each spectra and $W \times W$ represents the size of each band. The spectra and bands represent the spectral and spatial information of the sample, respectively. Motivated by the fact that the useful contents are invariant under the two domains, we learn representations coding such contents by pushing the matched bands and spectrum closer to each other with the SSCL framework.

2.2. Spatial-spectral contrastive learning

The SSCL is designed to learn powerful representations in four steps. First, we construct a pair of signals for every HSI sample to include the information in the two domains, respectively. Second, two embedding functions are built to learn the representations of different signals, respectively. Then, we employ two simple transform functions to project the representations to other spaces. Lastly, a contrastive loss is applied on the projections to guide the model training. The architecture of the SSCL is presented in Fig. 1, introduced next in details.

Table 1. Detailed settings of the 1D-CNN and 2D-CNN used as the embedding functions.

1D CNN		2D CNN	
Layer	Filters	Layer	Filters
Convolution	3, 64	Convolution	3×3, 64
Convolution Max pooling	[3, 128]×2 3, 128	Convolution Max pooling	[3×3, 128]×2 2×2, 128
Convolution Max pooling	[3, 256]×2 3, 256	Convolution Max pooling	[3×3, 256]×2 2×2, 256
Convolution Max pooling	[3, 512]×2 3, 512	Convolution Max pooling	[3×3, 512]×2 2×2, 512
Flatten Fully connected	N/A N/A	Flatten Fully connected	N/A N/A

Signal construction. Given a sample x^i, we construct a pair of spatial and spectral signals $(s^i_{\text{spa}}, s^i_{\text{spe}})$, which are regarded as a positive pair. In contrast, two signals of different samples are considered as a negative pair. The signal construction is crucial to the representation quality. A positive pair of signals are supposed to contain adequate spatial and spectral information of the object, respectively. Thus, the center spectrum is selected as the spectral signal with a size of C. Besides, we transform the sample by principal component analysis (PCA) and use the first component as the spatial signal with a size of $W \times W$. This solution is based on two considerations, covering abundant spatial information and refraining from mixing spectral information.

Embedding functions. We build two embedding functions $f_{\theta_{\text{spa}}}(\cdot)$ and $f_{\theta_{\text{spe}}}(\cdot)$ to learn the representations of the two signals, respectively. Our framework allows many options for the embedding functions. Considering the different size of the two signals, we adopt a two-dimensional convolutional neural network (2D-CNN) to learn the spatial representations $r_{\text{spa}} = f_{\theta_{\text{spa}}}(s_{\text{spa}}) \in \mathbb{R}^D$ and a 1D-CNN to obtain the spectral representations $r_{\text{spe}} = f_{\theta_{\text{spe}}}(s_{\text{spe}}) \in \mathbb{R}^D$. Note that each convolution layer in the two networks is followed by a batch-normalization layer and a ReLU activation function. The detailed settings of the two networks are presented in Table 1.

Transform functions. The representations are then projected onto other spaces with two non-linear transform functions $t_{\theta_{\text{spa}}}(\cdot)$ and $t_{\theta_{\text{spe}}}(\cdot)$. Specifically, we use two multilayer perceptrons (MLPs) to obtain the projections $p_{\text{spa}} = t_{\theta_{\text{spa}}}(s_{\text{spa}}) = W^2_{\text{spa}}\delta(W^1_{\text{spa}} r_{\text{spa}})$ and $p_{\text{spe}} = t_{\theta_{\text{spe}}}(s_{\text{spe}}) = W^2_{\text{spe}}\delta(W^1_{\text{spe}} r_{\text{spe}})$, where δ indicates the ReLU function, p_{spa} and $p_{\text{spe}} \in \mathbb{R}^L$. It is found that it is better to compute the contrastive loss on the projections than the representations directly [13].

Contrastive loss. A contrastive loss is defined on the acquired projections. We use the noise-contrastive estimation

(NCE) as the loss function [14]. Given a mini-batch of M samples, anchoring at the spatial signals, the loss can be computed as

$$L_{\text{NCE}}(s_{\text{spa}}, s_{\text{spe}}) = \sum_{i=1}^{M} -\log \frac{\exp(\Phi(p^i_{\text{spa}}, p^i_{\text{spe}})/\tau)}{\sum_{j=1}^{M} \exp(\Phi(p^i_{\text{spa}}, p^j_{\text{spe}})/\tau)}, \tag{1}$$

where τ is the temperature hyper-parameter and $\Phi(\cdot,\cdot)$ represents the cosine similarity between two vectors. The overall loss L_{NCE} is computed across anchoring both spatial and spectral signals. Minimizing this loss improves the prediction accuracy by encouraging the representations from the same samples to be similar while those of negative pairs dissimilar, which enables us to extract the contents shared between different domains. After the contrastive learning, the concatenation of spatial and spectral representation is used for the training of a linear classifier.

3. EXPERIMENTAL RESULTS

3.1. Dataset and implementation details

The effectiveness of the SSCL is verified on the widely-used Pavia University (PU) dataset. It is acquired over the campus of the Pavia University. After removing the noisy bands, 103 bands, each of which has a size of 610×340, are remained. 9 classes of different ground objects are included in the dataset. We select a 9×9 neighbourhood as the input sample. Thus, the patch size $W = 9$ and band number $C = 103$. 50% unlabeled samples are chose for the unsupervised training. During the classification evaluation, 5 labeled samples per class are randomly selected. We set the training epoch to 15, the learning rate to 3×10^4, the temperature hyper-parameter to 0.07 and the batch size M to 512. The Adam optimizer with $\beta_1 = 0.9$ is used to train the networks.

3.2. Results and analysis

In order to evaluate the representation quality, we conduct classification experiments on the PU dataset with 5 labels per class. Four unsupervised methods, namely, DAE [9], 3DCAE [10], WGAN-GP [12] and DMVL [15] are utilized for comparison. The comparison methods are reimplemented with PyTorch. The performance is measured by overall accuracy (OA), average accuracy (AA), class-specific accuracy and kappa coefficient (κ), where the OA represents the accuracy of all the test samples, AA is the average accuracy of each class and κ measures the agreement between the ground truth and classification results. For all the four metrics, higher value implies better performance.

As shown in Table 2, DAE obtains the lowest accuracy over the PU dataset, because it only considers spectral information for feature extraction. 3DCAE and WGAN-GP, which are based on 3D autoencoder and GAN, respectively,

Table 2. Classification results of different methods over the PU dataset. The best OA, AA and κ are in bold.

Class No.	DAE [9]	3DCAE [10]	WGAN-GP [12]	DMVL [15]	SSCL
1	51.47	58.60	44.08	77.04	79.29
2	67.82	63.69	75.97	77.19	93.99
3	25.71	71.41	92.74	93.56	88.00
4	96.42	95.71	98.09	91.86	99.60
5	99.93	100.00	100.00	100.00	100.00
6	65.14	76.20	81.43	73.99	87.29
7	97.59	98.80	99.55	98.27	95.11
8	81.59	90.39	54.21	80.12	88.65
9	99.26	100.00	99.79	99.79	95.99
OA(%)	69.04	72.76	74.24	80.86	**90.79**
AA(%)	76.10	83.87	82.87	87.98	**91.99**
κ(%)	61.71	67.00	68.05	75.87	**88.08**

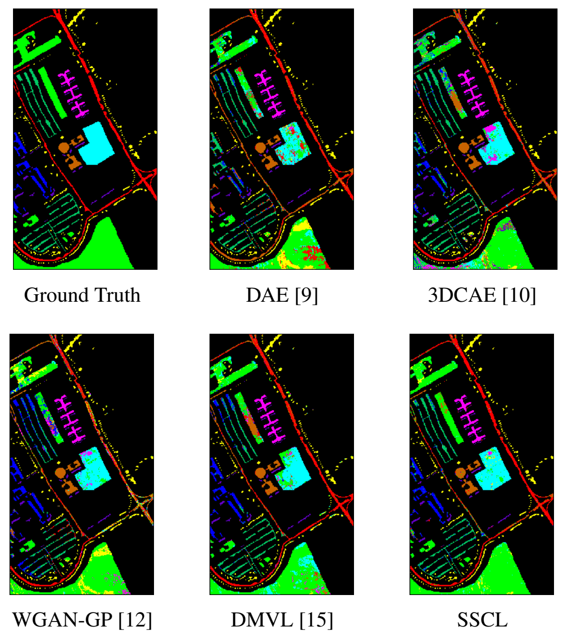

Fig. 2. Classification maps of each method over the PU dataset.

also give quite poor performance in terms of various metrics. It demonstrates that compressing the information of the HSI samples undiscriminately is not effective for extracting powerful features. Compared to them, DMVL presents better performance, since it extracts more useful contents by conducting contrastive learning on different bands. However, many nuisance information are also shared between different bands, which lowers the representation quality greatly. In contrast, our method extracting features by contrasting the spectral and spatial domains outperforms all the comparison methods significantly, which demonstrates the superiority of capturing the domain-invariant information over other methods for discovering the desired semantic features and discarding the less useful contents. For further comparison, the classification maps of different methods and the ground truth over the PU dataset are presented in Fig. 2. As can be seen, the SSCL exhibits better classification result compared to the state-of-the-art methods, which also verifies the quantitative assessment results.

4. CONCLUSION

In this work, a spatial-spectral contrastive learning framework is presented to learn powerful representations of HSIs. It is motivated by the observation that the contents useful for classification are invariant under the spectral and spatial domains while the less useful contents are not. Thus, our framework is designed to learn representations that capture the domain-invariant information. The results of classification experiment verifies the discriminative ability of our representations. Our work brings new opportunities for the unsupervised representation learning of HSIs by exploring the information shared between the two domains.

5. REFERENCES

[1] Thais Andressa Carrino, Alvaro Penteado Crósta, Catarina Labouré Bemfica Toledo, and Adalene Moreira Silva, "Hyperspectral remote sensing applied to mineral exploration in southern Peru: A multiple data integration approach in the Chapi Chiara gold prospect," *International Journal of Applied Earth Observation and Geoinformation*, vol. 64, pp. 287–300, February 2018.

[2] Foudan Salem, Menas Kafatos, Tarek El-Ghazawi, Richard Gomez, and Ruixin Yang, "Hyperspectral image analysis for oil spill detection," in *Summaries of NASA/JPL Airborne Earth Science Workshop, Pasadena, CA*, 2001, pp. 5–9.

[3] Chunsen Zhang and Yiwei Zheng, "Hyperspectral remote sensing image classification based on combined SVM and LDA," in *Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications V*, November 2014, vol. 9263, p. 92632P.

[4] Nan Meng, Hayden Kwok-Hay So, Xing Sun, and Edmund Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 873–886, March 2021.

[5] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, July 2017.

[6] Peiyan Guan and Edmund Y Lam, "Multistage dual-attention guided fusion network for hyperspectral pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, October 2021.

[7] Shiqi Yu, Sen Jia, and Chunyan Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, January 2017.

[8] Bin Zhang, Lin Zhao, and Xiaoli Zhang, "Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images," *Remote Sensing of Environment*, vol. 247, pp. 111938, September 2020.

[9] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, June 2014.

[10] Shaohui Mei, Jingyu Ji, Yunhao Geng, Zhi Zhang, Xu Li, and Qian Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6808–6820, September 2019.

[11] Mingyang Zhang, Maoguo Gong, Yishun Mao, Jun Li, and Yue Wu, "Unsupervised feature extraction in hyperspectral images based on wasserstein generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2669–2688, May 2019.

[12] Qiaoqiao Sun and Salah Bourennane, "Hyperspectral image classification with unsupervised feature extraction," *Remote Sensing Letters*, vol. 11, no. 5, pp. 475–484, February 2020.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *2020 International Conference on Machine Learning*, 2020, vol. 119, pp. 1597–1607.

[14] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *International Conference on Artificial Intelligence and Statistics*, May 2010, vol. 9, pp. 297–304.

[15] Bing Liu, Anzhu Yu, Xuchu Yu, Ruirui Wang, Kuiliang Gao, and Wenyue Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, September 2021.