MIXED VISION TRANSFORMER AND GRAPH NEURAL NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Paul-Beniamin Ilioaica, Anamaria Radoi

University Politehnica of Bucharest

ABSTRACT

Hyperspectral images contain information across the electromagnetic spectrum for every pixel in the remote sensing image. This enables the identification of various materials and objects with different spectral signatures at a high level of precision. In this paper, we propose an end-to-end framework for hyperspectral image classification based on Graph Convolutional Networks (GCN) and Vision Transformers (ViT) that extract both neighborhood information and spatial relational information. We evaluate the performance of the proposed methodology on two benchmark hyperspectral datasets, namely Indian Pines and Pavia University, and compare the achieved performance with state of the art.

***Index Terms*—** Hyperspectral image classification, Vision transformers, Graph convolutional neural networks.

1. INTRODUCTION

Hyperspectral (HS) images carry a large volume of spectral information that allows an improved recognition of materials and land cover classes, being, thus, a powerful tool for image semantic segmentation. With the emergence of deep learning techniques, the interpretation of HS images has received a strong impulse in the recent years. Several methods based on convolutional neural networks (CNNs), including contextual deep CNN exploiting both spatial and spectral features [1] and 3-D CNN [2], have been proposed. In addition, in order to reduce the influence of interfering pixels that have different spectral signatures at the edge of each land-cover area, a spectral-spatial attention network has been integrated into a spectral spatial network based on 3-D CNNs [3].

Although CNNs are extremely efficient in tasks related to image classification, object detection and semantic segmentation, these architectures analyze patches of data on a fixed grid, with possibly stride convolutions and pooling, focusing more on spatial content information rather than learning more complex dependencies, either from a spatial or spectral perspective. In this regard, CNNs are not able to exploit relations between non-neighborhood pixels. In an attempt to convert HS data into irregular domains from regular grids, a localized graph convolutional filtering based on spectral graph theory has been applied to HS images [4], where the graph convolutional layer is used to learn the spatially local graph representation and to identify localized topological patterns of the graph nodes.

Traditional GCN methods rely on the construction of the adjacency matrix, which increases the computational burden of the algorithms, especially in the case of large volumes of data. In order to be able to train large-scale GCN using a mini-batch approach, Hong *et al.* proposed miniGCN, which combines the power of CNNs and GCNs to extract and fuse HS features in an end-to-end trainable network [5]. Moreover, three fusion strategies have been considered in [5], namely additive, element-wise multiplicative, and concatenation fusion strategies.

Recently, Hong *et al.* proposed SpectralFormer [6], which aims at learning locally spectral representations from multiple neighboring bands rather than single band. Without other convolutional and recurrent layers, transformers show great potential in extracting detailed spectral representations, reaching competitive results compared to CNN, GNN or recurrent neural network (RNN)-based approaches.

Considering the advantages of GCNs, e.g. extension to irregular domains, modeling topological relations, and improvement of class boundaries, in this paper, we propose a novel method which extends the benefits of GCNs through the inclusion of vision transformers layers. This enables the extraction of local knowledge through an attention mechanism in addition to aggregating unstructured data by means of GCN.

2. PROPOSED METHOD

The proposed methodology for hyperspectral image semantic segmentation, extracts both structured and unstructured data and makes use of mixed information retrieved by Graph Convolutional Networks (GCN) and Vision Transformers (ViT). The main advantage of different network typologies is represented by the ability to extract heterogeneous information from data, such as neighbourhood information through the ViT architecture and spatial relational information through

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number PN-III-P1-1.1-PD-2019-0843 (MDM-SITS), within PNCDI III. E-mail: anamaria.radoi@upb.ro

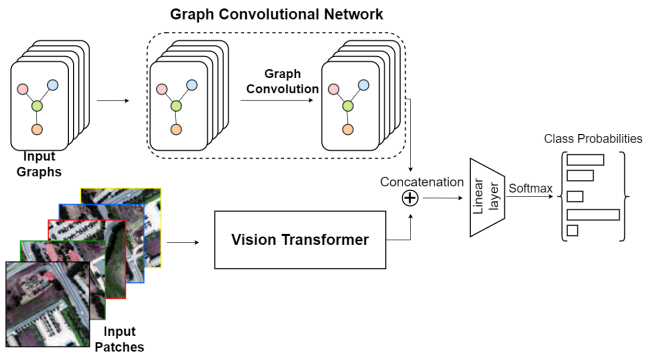

Fig. 1. Model architecture

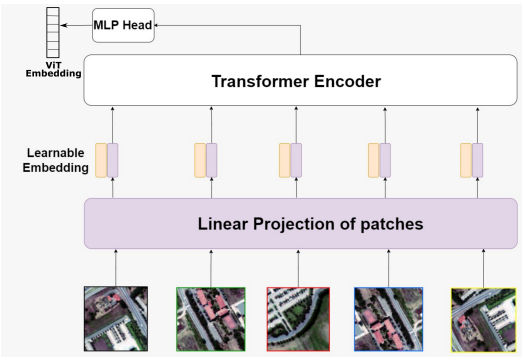

Fig. 2. ViT module

the graph representation. The pixel-wise concatenation of GCN and ViT-based representations is fed into a simple linear layer, followed by a Softmax layer which transforms the logits into a probability distribution, in order to predict the semantic class. The scheme of the proposed model architecture is presented in Fig. 2, whereas each block components, i.e. GCN and ViT-based modules, will be detailed in the following sections.

2.1. Graph Neural Networks

In general, unstructured data is hard to manipulate and even harder to process. In this context, graph representations have shown an immense potential and managed to outperform traditional approaches based on CNNs. First introduced in [7], GCN are an emerging methodology that is reinvigorating numerous machine learning-based approaches by offering a natural framework to generalize the notion of convolution. This is achieved by learning node embeddings through multilayer local aggregation.

Considering a graph $G = (V, E)$, with V, E representing the vertex and edge sets, respectively, a GCN receives, as input, the feature matrix and the adjacency matrix of the graph, denoted by A. If N is the number of vertexes, the element $A_{i,j}$ of the adjacency matrix provides the relationship between the vertexes V_i and V_j, which can be computed using a radial basis function:

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \tag{1}$$

The core of Graph Neural Networks (GNN) is the message passing methodology. The message passing methodology spreads each node feature to the neighborhood of nodes using trainable weights. There are multiple ways to implement the trainable weights, such as shareable with respect to the distance between the nodes [8], shareable to the connected node features [9] or to node/edge features. In GCN, the network is composed of layers, denoted by $H^{(l-1)}$, that take in the node representation of the previous layer $H^{(l-1)}$ and compute H^l, with $H(0)$ being the initial patch. The general form of a GCN can be written as:

$$H^{(l+1)} = \sigma\left(\sum_s C^{(s)} H^{(l)} W^{(l,s)}\right) \tag{2}$$

where $C^{(s)}$ is the s^{th} convolution support which defines the node feature propagation to neighbour nodes, $W^{(l,s)}$ is the trainable matrix for the l^{th} layer. It is worth mentioning that various GCN architectures differ in the design of the convolution support $C^{(s)}$ [8, 9].

In the hyperspectral image classification setting considered in our approach, the vertexes V of the graph G hold the spectral information contained within the hyper-spectral pixels, whereas the edges E represent the similarity between different pixels. The GCN-based module presented in Fig. 2 is responsible for graph processing by message passing. Considering graph convolutions, the message passing is performed via a graph convolution layer, followed by a batch normalization layer, which is used to increase stability and decrease variance in training, and an activation function. In general, a dropout layer is applied to avoid overfitting and increase generalization.

2.2. Vision Transformers

With simple architectures based on attention mechanisms, transformers show superior precision and significant less time for training, outperforming deep convolutional neural networks [10]. The Transformer itself is a sequence-to-sequence model that is composed of an encoder and decoder, e.g. standard ViT interprets the image as a sequence of patches and employs a standard Transformer encoder architecture, which has been initially proposed for natural language processing tasks [11]. Similarly, the image is split into patches, which are linearly embedded. The concatenation of the patches and the embeddings is fed into the transformer encoder module and is treated in the same way as tokens (words). The encoder module contains the self-attention module and position-wise

feed-forward network, whereas the decoder module additionally adds cross-attention modules between the position-wise feed-forward networks and the multi-head self-attention.

As shown in Fig. 1, the inputs of the ViT module are the image patches centered on the pixel of interest, which are flattened, concatenated with extra learnable class embeddings, and passed through the ViT's encoder module. As mentioned, the key element of ViT is the attention module, which is a function that maps a query and a set of key-value pairs to an output though a multi-head architecture:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

The query and key-value pairs for N input tokens stored in $X \in \mathbb{R}^{N\times d}$ (i.e., d is the token dimension) are given by $Q = X \cdot W_q$, $K = X \cdot W_k$, $V = X \cdot W_v$, where $W_q \in \mathbb{R}^{d\times d_q}$, $W_k \in \mathbb{R}^{d\times d_k}$, $W_v \in \mathbb{R}^{d\times d_v}$, h being the number of heads and $d_q = d_k = d_v = d/h$. Each projection learns a different mapping with all being averaged in the final steps.

3. EXPERIMENTS

In order to assess the performance of the proposed method, we used two benchmark datasets for hyperspectral image classification, namely Pavia University and Indian Pines datasets. The Pavia University dataset consists of 612×340 pixels (9 classes), which are characterized by 113 spectral bands from 0.43 to 0.86 μm and a spatial resolution of 1.3 meters. The Indian Pines dataset consists of a multi-spectral image of 145×145 pixels (16 classes) with 224 spectral bands ranging from 0.4 to 2.5 μm and a spatial resolution of 20 meters. In the following, for each pixel, we consider a patch of 7×7 pixels around it. We used a 2-layers GCN architecture, with 64-dimensional internal graph node embeddings and a dropout of 0.5, whereas the number of heads in the attention mechanism of ViT was fixed to 8. The dimension of the embeddings generated by both ViT and GCN modules is equal to the number of classes for each dataset, whereas the last linear layer performs the classification on the concatenated embeddings.

For both datasets, we split the pixels into two disjoint sets with 30% for training and 70% for testing by random sampling with a uniform distribution among class labels. We propose using a transductive method for data splitting, which ensures that no data leakage occurs between training and testing pixels. Training and testing data generation is done in separate steps on distinctive sets of sampled pixels. The generation of training data is done according to the following steps:

(i) sample v pixels with replacement,

(ii) compute A, V, E for the graph, and

(iii) store the patches centered on the sampled pixels.

Through sampling with replacement, we are able to perform a simple augmentation and generate a sufficient number of training graphs, considering that the training dataset is rather small. We found that 190 samples, i.e. constructing graphs with 190 vertexes, provided the best performance, while not being too memory intensive. For both cases, 1000 training graphs were sampled proving that this augmentation technique offers strong performance benefits for small datasets. For the testing set, exhaustive sampling without replacement is done to generate the test graphs.

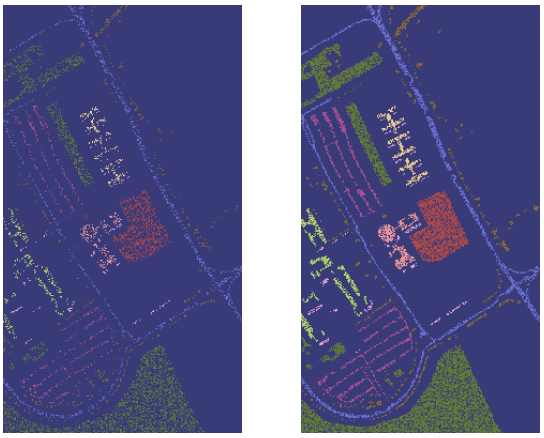

Fig. 3. Spatial distribution of training set (first column) and testing set (second column).

We compared our combined GCN and ViT-based approach with other methods in the literature, in terms of overall accuracy (OA) and Kappa coefficient (κ). The results are shown in Table 1. The proposed algorithm achieved an overall accuracy of 99.10% and a Kappa coefficient of 0.98, which is similar to the best performing approaches for both the Pavia University [2] and the Indian Pines datasets [3]. Moreover, the normalized confusion matrices are presented for both datasets in Fig. 4 and Fig. 5. For a visual evaluation of the results, we show the predicted and ground truth maps for both datasets in Fig. 6 and Fig. 7.

4. CONCLUSION

We proposed a new approach for the semantic segmentation of HS images based on a mixed GCN and ViT framework. Combining knowledge extracted with powerful graph neural networks and high-performing transformer-based architectures proved successful for HS image classification, reaching an overall accuracy of 99.10 % on Pavia University and 95.53% on Indian Pines datasets. Using simple, but powerful architectures, the results are in line with the best performing architectures in the literature. Furthermore, we proposed a new method of generating data for the train dataset by random sampling with replacement, which also helps augment training data that represents a small fraction of all.

Table 1. Performance results and comparisons with other approaches (- indicates that the value was not provided).

No.	Method	OA[%]	κ
Pavia University	Proposed method	99.10	0.98
	SpectralFormer [6]	91.07	0.88
	miniGCN [5]	79.79	0.73
	FuNet-C [5]	92.20	0.89
	Deep CNN [1]	0.94	-
	SSAN [3]	98.02	0.97
	3D-CNN-LR [2]	99.54	0.99
Indian Pines	Proposed method	95.53	0.94
	SpectralFormer [6]	81.76	0.79
	miniGCN [5]	75.11	0.71
	FuNet-C [5]	79.89	0.77
	Deep CNN [1]	0.92	-
	SSAN [3]	95.49	0.94
	3D-CNN-LR [2]	97.56	0.97

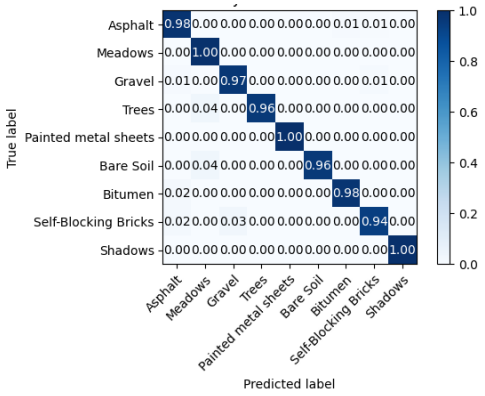

Fig. 4. Confusion matrix for Pavia University.

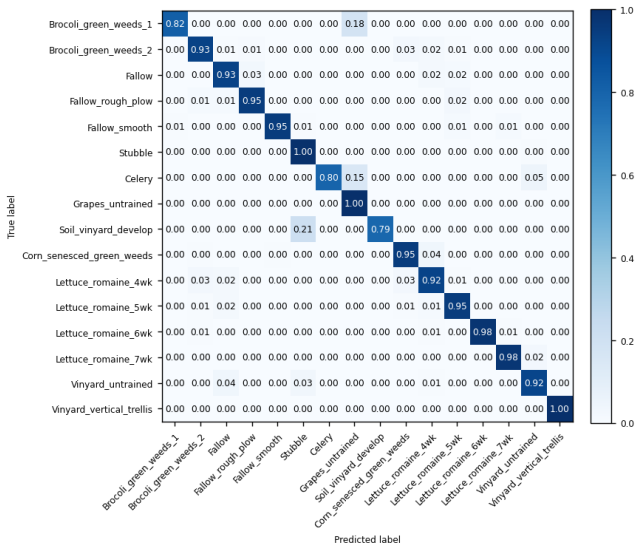

Fig. 5. Confusion matrix for Indian Pines.

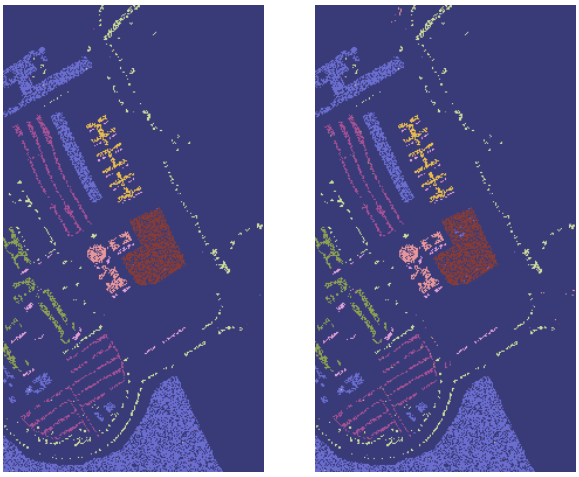

Fig. 6. Ground truth (left) and predicted map (right) for Pavia University.

5. REFERENCES

[1] H. Lee and H. Kwon, "Contextual deep cnn based hyperspectral classification," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 3322–3325.

[2] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[3] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3232–3245, 2020.

[4] S. Pu, Y. Wu, X. Sun, and X. Sun, "Hyperspectral image classification with localized graph convolutional filtering," *Remote Sensing*, vol. 13, no. 3, 2021.

[5] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.

[6] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *CoRR*, vol. abs/2107.02988, 2021.

[7] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *CoRR*, vol. abs/1606.09375, 2016.

[9] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin,

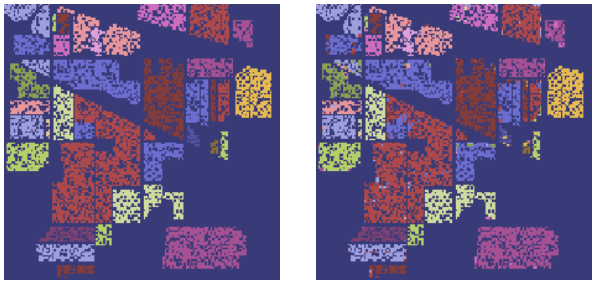

Fig. 7. Ground truth (left) and predicted map (right) for Indian Pines.

"Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.