ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413664

UNSUPERVISED STACKED CAPSULE AUTOENCODER FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Erting Pan[1], *Yong Ma*[1,2], *Xiaoguang Mei*[1,2], *Fan Fan*[1,2,*], *Jiayi Ma*[1,2]

[1]School of Electronic Information,Wuhan university, Wuhan, 430072, China,
[2]Institute of Aerospace Science and Technology, Wuhan university, Wuhan, 430072, China

ABSTRACT

Since CapsNet [1] shattered all previous records of algorithms for image recognition, the capsule's conception has attracted bright attention. It interprets an object by the geometrical arrangement of parts. We think it can be transferred to hyperspectral images. In a hyperspectral data cube, each pixel spectrum can be regarded as a continuous curve representing its inherent properties. In the spatial domain, there are various spatial distributions in different positionsand there is usually a specific structural relationship between adjacently distributed categories. Based on HSI data's aforementioned structural characteristics, combined with the stacked capsule autoencoder, we propose our model to achieve an unsupervised HSI classification. In our model, the ConvLSTM is employed to discover part capsules of HSI, and we utilize Set Transformer to encode relations among all parts and indicate object capsules. The decoders of both phases use Gaussian mixture models to reconstruct specific information. Experimental results of the Pavia Center dataset show the exceptional of our model.

Index Terms— hypersepctral classification, unsupervised learning, capsule network, stacked autoencoder

1. INTRODUCTION

Hyperspectral image (HSI) has received extensive attention in the field of remote sensing with the advantage of owing rich spectral and spatial information. In particular, its high spectral resolution can reflect the different characteristics in detail of different objects in the spectrum and be applied to various industries, such as precision agriculture, urban planning, environmental monitoring, and geological prospecting [2, 3].

HSI classification is the most fundamental issue to achieve the applications as mentioned above. It refers to assign a specific label to each pixel in HSI. However, the essential characteristics of HSI, including high dimensionality, enormous structure complexity, massive information redundancy between adjacent bands, make HSI classification a very challenging task [4, 5]. A standard solution to this critical issue is focused on how to extract robust and discriminative features. Compared to traditional methods that rely on manual features [6, 7], the deep learning algorithm has a hierarchical structure to extract high-level features. It has become a useful and competitive tool for many tasks [8]. A vast amount of work utilizing deep learning methods for HSI classification has been proposed recently, especially using a convolutional neural network (CNN) and its various variants [9, 10, 11].

Despite many deep learning methods that have shown their superiority for HSI classification, some problems still have adverse effects on the performance. The issue of insufficient labeled samples comes first. It is limited by expensive cost in collecting and labeling hyperspectral data, only a small part of pixels are labeled in existing public HSI datasets while CNN requires sufficient labeled data for training [12]. Second, due to the abundant spectral information and complex spatial distribution of HSI data, the outstanding performance of HSI classification usually relies on a complex network structure heavily, such as 3D CNN [13] and a diverse combination of CNN and other powerful techniques [14, 15]. It suffers from high computational cost. Thirdly, CNN usually consists of stacked convolution and pooling layers. One advantage of the pooling layer is that it brings local translation invariance, which addresses the sensitivity of the feature location. It is achieved by summarizing the presence of features, which also reduces the size of feature maps. Nevertheless, at the same time, it causes an unavoidable loss of valuable information [16].

To solve the limitations exhibited by the pooling layer, in 2017, Sabour *et al.* [1] proposed Capsule Networks based on capsules and dynamic routing. The extraordinary of the capsule is that it could preserve hierarchical pose relationships between object parts. Unlike CNN, which extracts scalar features, the capsule's output incorporates a set of vector neurons representing the objects' various attributes such as presence, position, scale, orientation, texture, and beyond. Moreover, the capsule integrates close relationships between objects and represents numerically. Some recent works which employed this architecture have shown its capabilities for HSI classification. For example, Paoletti *et al.* [17] designed a spectral-spatial CapsNet for HSI classification, and Deng *et al.* [18] proposed a model called HSI-CapsNet, which only employ a

Fan Fan is the corresponding author. This research was funded by National Natural Science Foundation of China under grant No. 61903279.

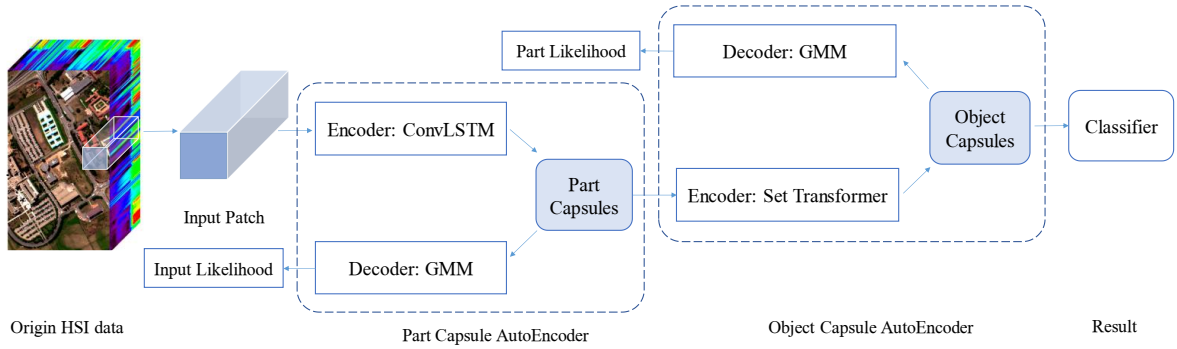

Fig. 1: Illustration of stacked capsule autoencoder for HSI classification. It contains a part capsule autoencoder and a object capsule autoencoder.

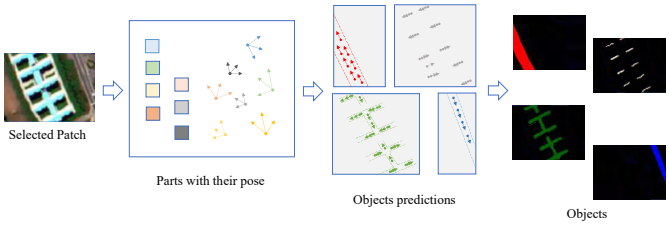

Fig. 2: The modeling process of our method. The parts with poses are learned in the part capsule autoencoder, and the object predictions are derived from the object capsule autoencoder.

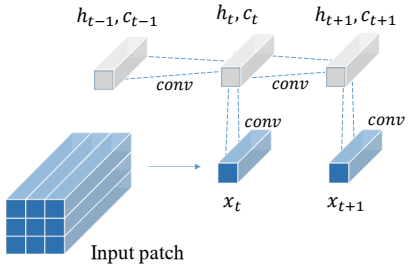

Fig. 3: Illustration of ConvLSTM network, which could extract spectral and spatial features.

limited number of training samples.

Lately, Kosiorek *et al.* [16] proposed an unsupervised version of a capsule network in which highly structured decoder networks train both an encoder network that can segment images into parts and their poses and an encoder network that can compose these parts into coherent wholes. Inspired by this, associate the spectral and spatial characteristics of HSI data, we utilize the capsule conception to model HSI and understand it with the association between parts and objects. We pertinently advance an unsupervised stacked autoencoder for HSI classification.

Our work's merits are as follows: (1) Aiming at the structure of HSI data, we employ ConvLSTM to extract spectral-spatial features and derive part capsules with reliable parameters. (2) In our model, the object capsules represent different classes of an HSI. The object-viewer relations explain different spatial distributions, and the object-part correlations describe the similarity of spectral-spatial features between adjacent bands.

2. METHODOLOGY

Based on the stacked capsule autoencoder and HSI data's structural characteristics, we design our model to achieve an unsupervised HSI classification, as shown in Fig. 1. The whole architecture can be described as two phases, including the part capsule autoencoder and the object capsule autoencoder. The modeling process for HSI classification is illustrated in Fig. 2. In this section, we give a detailed introduction to our method.

2.1. Part Capsule Autoencoder

The first phase is the part capsule autoencoder, which encodes the input patch to obtain part capsules with its unique features, poses, and presence probabilities, and decodes each part to templates to reconstruct the input patch by affine-transformed templates.

Instead of the original encoder CNN with attention, considering the structural characteristics of HSI, we utilize ConvLSTM [19] as the encoder to acquire part capsules and infer their unique features z_M, presence probabilities α_M, and pose e_M. The design of ConvLSTM is shown in Fig. 3. Specifically, the main difference between ConvLSTM and LSTM is that the former uses a convolutional structure to replace the forward transmission of the input and states. In detail, the input of each ConvLSTM cell is set to each pixel with the whole spectral vector in the input patch, and adjacent pixels input the model following timesteps. It has the advantage of extracting spatial and spectral features simultaneously for H-

SI data and has a positive effect on inferring reliable part capsule parameters. Moreover, the pose of part capsule is defined as six-dimensional, including two rotations, two translations, scale, and shear.

Let $x \in [0,1]^{h\times w\times b}$ be the input patch of HSI data, b represents the number of spectral bands. In particular, since different spectral bands of an HSI have the same spatial distribution characteristics, we consider reconstructing the input in the spatial domain. The maximum number of part capsules is limited to M, and the template is set to $T_M \rightarrow [0,1]^{h\times w\times 1}$. We reconstruct the input with a spatial Gaussian Mixture Model (GMM).In our model, the centers of isotropic Gaussian components areare the transformed templates' pixels, and the variance is set as a constant. The input likelihood is given by,

$$p(x) = \prod_{i,j}\sum_{m=1}^{M} p_{m,i,j}\mathcal{N}(x_{i,j}|\hat{T}_{m,i,j};\sigma_x^2), \tag{1}$$

where $\hat{T}_{m,i,j}$ represents the pixel in the affine-transfromed templates, and the mixing probabilities $p_{m,i,j}$ is is proportional to product of $\hat{T}_{m,i,j}$ and the part capsules' presence probability α_M.

The target of training this phase is to extract reliable parts, and it also results in learning templates for parts. Its object function is to maximize the input likelihood of Eq. 1.

2.2. Object Capsule Autoencoder

For the task of HSI classification, we define the object capsules in HSI as different classes. Based on identified parts and their parameters, we can discover how parts compose objects in the spatial domain at the object capsule autoencoder phase.

In this phase, the object capsule autoencoder encodes part capsules to acquire their inner relations and reconstruct the part pose with a separate mixture of each part's predictions using object capsules. The encoder of this phase is the Set Transformer [20], which is an attention-based set-input neural network architecture. As part capsules obtained from the former phase are order-independent, Set Transformer is a suitable choice for modeling complicated interactions among all parts.

Instead of employing all the part capsule parameters directly, we concatenate the flattened templates, unique features, and poses as its input. Besides, considering each object is composed of $N(N \leq M)$ part candidates, we feed-in the presence probabilities of the part capsule to the encoder separately. The output of Set Transformer is K object capsules with their parameters, including special features c_k, presences probabilities β_k, and an object-viewer relation matrix OV_k, which represents the different spatial distribution of the same class in an HSI.

Table 1: Summary of the parameter settings.

Name	Value
size of spatial patch	27×27
learning rate	$1e-4$
number of part capsules	24
number of object capsules	10
dimension of part features	16
dimension of object features	16

After that, in order to discover which part capsules composed the object capsule, we decode the feature of the object capsule to obtain the parameters related to part candidates. A separate multilayer perceptron (MLP) is employed to acquire the conditional probability $\beta_{k,m}$, an associated scalar standard deviation $\lambda_{k,m}$, and an object-part relationship matrix $OP_{k,m}$. Additionally, the candidate predictions $\mu_{k,m}$ are the product of the object-viewer relation matrix OV_k and the object-part relation matrix $OP_{k,m}$, formally, $\mu_{k,m} = OV_k \cdot OP_{k,m}$, which represents intra-class correlations of the same class in an HSI.

On this basis, we reconstruct the part pose with an independent mixture of predictions from object capsules. To be specific, the likelihood of part capsules is given by,

$$p(z_m,\alpha_m) = \prod_{m=1}^{M}[\sum_{k=1}^{K}\frac{\beta_k\beta_{k,m}}{\sum_i\beta_i\sum_j\beta_{i,j}}\mathcal{N}(z_m|\mu_{k,m},\lambda_{k,m})]^{\alpha_m}, \tag{2}$$

where $u_{k,m}$ and $\lambda_{k,m}$ are the centers and standard deviations of the isotropic Gaussian components.

Training this phase is to study the interactions of different parts, which indicates the intra-class correlation and further spatial distribution of the same class in HSI classification. The object function of this phase is maximizing the part pose likelihood of Eq. 2.

3. EXPERIMENTS

We train and test our method on public hyperspectral image classification dataset, namely, the Pavia Center dataset. It is collected by ROSIS sensor, and it contains nine land cover classes of urban areas. Pavia University dataset has a spatial size of 1096×715 with 102 spectral bands.

To demonstrate the superiority and effectiveness of our method, we compare it with the traditional unsupervised classification method KNN, typical deep learning models RNN and 2DCNN, and ConvLSTM, which is also used for discovering part capsules in our model. We split the training and testing dataset with a random sampling strategy, and the sampling percentage is set to 0.1. For a fair comparison, we utilize the same training and testing sets for all methods, and all algorithms are executed ten times. Besides, we employ standard classification metrics, including overall accuracy (OA),

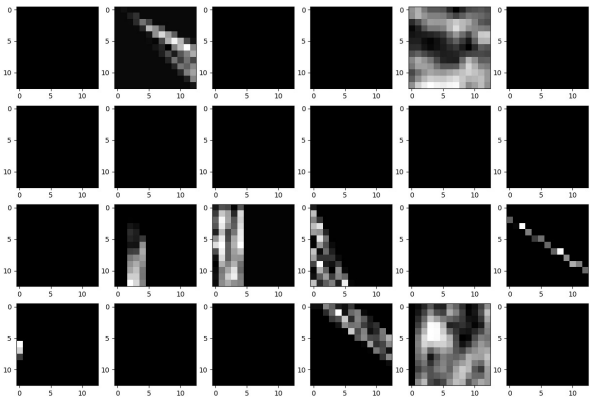

Fig. 4: Visualization of templates that learned through part capsule autoencoder, which is trained independently with five epochs, and the number of part capsules is set as 24.

Table 2: Classification performance of different methods for the Pavia Center dataset. Bold indicates the best result.

Label	KNN	RNN	2DCNN	ConvLSTM	our method
OA	88.29	90.75	92.75	93.47	**96.39**
AA	89.14	93.88	87.33	96.25	**97.16**
Kappa	82.59	87.66	89.94	90.16	**95.88**

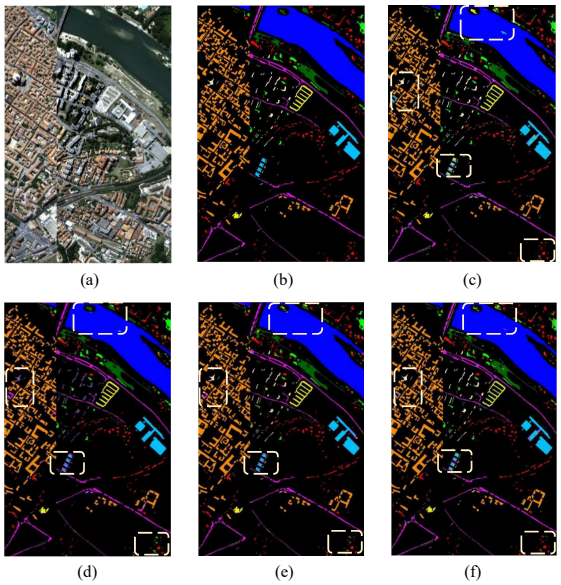

Fig. 5: Comparison of classification maps of different methods. (a) false-color image, (b) groundtruth, (c) RNN, (d) 2DCNN, (e) ConvLSTM, (f) our method. The white dotted frames have marked the region with obvious differences in classification maps.

average accuracy (AA), and kappa. All the experiments are implemented with an Intel Xeon Gold 5117, 2.00GHz, and an NVIDIA RTX 2080Ti GPU. Some settings of vital parameters related to our model are list in Table 1.

Fig. 4 shows templates learned through part capsule autoencoder, and it is just the result of part capsule autoencoder after independent training in five epochs. The initial settings of templates are randomly initialized into a fixed size. In the part capsule autoencoder, the part capsule's characteristics, including unique features, poses, and presence probabilities, are learned during training, and the transformed templates are gradually optimized by reconstructing the input patch through GMM. Besides, to be specific, in our method, the part capsule autoencoder and object capsule autoencoder is co-training for more well-balanced classification performance.

The Pavia Center dataset's experimental results are shown in Tabel 2, and Fig. 5 represents their classification maps. Both the quantitative and qualitative results exhibit the best performance among all compared methods. It indicates that our proposed method is effective in HSI classification. The traditional method KNN demonstrate poor performance. Deep learning methods, RNN [21], and 2DCNN [22] are competitive because of their discriminative features. The former has better overall accuracy performance, while the latter has better average accuracy with a more uniform classification map. Compared to RNN and 2DCNN, ConvLSTM also shows its extraordinary because it could simultaneously extract spectral and spatial features. In our method, ConvLSTM is used to discover the parts with their specific parameters. After that, in order to find object-part relation and object-viewer relation in HSI, features obtained from part capsule autoencoder are encoded again in object capsule autoencoder. In this way, our method gains a preferable and exceptional result among all compared methods.

4. CONCLUSION

In this study, an unsupervised stacked capsule autoencoder is proposed for the HSI classification task. Our method is designed based on the capsule concept, which can express features such as posture, direction, and positional relationship in the spatial domain. HSI can use capsules to describe the rich structural features inside. In spatial, due to the interaction of objects, adjacent pixels' spectral characteristics usually have a certain similarity. Besides, the same class of objects usually has different distributions in different positions. Therefore, the capsule is very suitable to use object-viewer and object-part correlations to model HSI.

In our model, the ConvLSTM is employed to discover part capsules of HSI, and we utilize Set Transformer to encode relations among all parts and indicate object capsules. The decoders of both phases use Gaussian mixture models to reconstruct specific information. The experimental results of the Pavia Center dataset show the superior of this model. In light of this study, we believe the capsule concept has broad prospects in HSI classification, and further efforts should be devoted to developing a more concise model soon.

5. REFERENCES

[1] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[2] He Lin, Jun Li, Chenying Liu, and Shutao Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–19, 2018.

[3] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jn Atli Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[4] Fan Fan, Yong Ma, Chang Li, Xiaoguang Mei, Jun Huang, and Jiayi Ma, "Hyperspectral image denoising with superpixel segmentation and low-rank representation," *Information Sciences*, vol. 397, pp. 48–68, 2017.

[5] Pedram Ghamisi, Naoto Yokoya, Jun Li, Wenzhi Liao, Sicong Liu, Javier Plaza, Behnood Rasti, and Antonio Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.

[6] Jun Li, Prashanth Reddy Marpu, Antonio Plaza, Jose M Bioucas-Dias, and Jon Atli Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, 2013.

[7] Yong Ma, Yuanshu Zhang, Xiaoguang Mei, Xiaobing Dai, and Jiayi Ma, "Multifeature-based discriminative label consistent k-svd for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4995–5008, 2019.

[8] Liangpei Zhang, Lefei Zhang, and Bo Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[9] Hyungtae Lee and Heesung Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017.

[10] Leyuan Fang, Guangyun Liu, Shutao Li, Pedram Ghamisi, and Jn Atli Benediktsson, "Hyperspectral image classification with squeeze multibias network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1291–1301, 2018.

[11] ME Paoletti, JM Haut, J Plaza, and A Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 279–317, 2019.

[12] Junjun Jiang, Jiayi Ma, Zheng Wang, Chen Chen, and Xianming Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 851–865, Feb 2019.

[13] Ying Li, Haokui Zhang, and Qiang Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, vol. 9, no. 1, pp. 67, 2017.

[14] Xiangyong Cao, Feng Zhou, Lin Xu, Deyu Meng, Zongben Xu, and John Paisley, "Hyperspectral image classification with markov random fields and a convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[15] Xiaoguang Mei, Erting Pan, Yong Ma, Xiaobing Dai, Jun Huang, Fan Fan, Qinglei Du, Hong Zheng, and Jiayi Ma, "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 8, 2019.

[16] Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton, "Stacked capsule autoencoders," in *Advances in Neural Information Processing Systems*, 2019, pp. 15512–15522.

[17] Mercedes E Paoletti, Juan Mario Haut, Ruben Fernandez-Beltran, Javier Plaza, Antonio Plaza, Jun Li, and Filiberto Pla, "Capsule networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2145–2160, 2018.

[18] Fei Deng, Shengliang Pu, Xuehong Chen, Yusheng Shi, Ting Yuan, and Shengyan Pu, "Hyperspectral image classification with capsule network using limited training samples," *Sensors*, vol. 18, no. 9, pp. 3153, 2018.

[19] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[20] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3744–3753.

[21] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[22] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram 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.