AI-NET: ATTENTION INCEPTION NEURAL NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Zhitong Xiong[1], *Yuan Yuan*[1], *Qi Wang*[1,2*]

[1] School of Computer Science and Center for OPTical IMagery Analysis and Learning,
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.
[2] Unmanned System Research Institute,
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

ABSTRACT

Recently, Deep learning methods have dominated many filed-s thanks to its powerful discriminative feature learning ability. While for hyperspectral images (HSI) analysis, these deep neural networks methods suffer from overfitting as the number of labeled training samples are limited. Thus more efficient neural network architecture should be designed to improve the performance of HSI classification task. In this paper, a novel attention inception module is introduced to extract features dynamically from multi-resolution convolutional flters. The AI-NET constructed by stacking the proposed attention inception module can adaptively adjust the network architecture by dynamically routing between the attention inception modules. By exploiting different spatial size convolutional filters and dynamic CNN architecture, more representative feature can be learned with limited training samples. Extensive experimental results have shown that the proposed method can adaptively adjust the network atchitecture and obtain better classification performance.

***Index Terms*—** Hyperspectral image classification, Attention model, Inception model, Dynamic routing, Deep learning

1. INTRODUCTION

Hyperspectral imaging data contains rich spectral and spatial information which are useful for a variety of earth remote sensing applications. Among them, pixel-wise hyperspectral image classification provides the spatial distribution of the land cover classes, and it is pivotal for the investigation of environmental processes. While great effort has been made for the research of classification algorithms, there still some issues to cope with, for example, the high dimensionality of pixels, complicated spatial distribution of different land cover classes and the limited training samples.

As aforementioned, hyperspectral images provide richer information by the hundreds of spetral channels. However, when the HSI spectral bands is more than the training samples, the classification accuracy decreases dramatically with the increase of data dimensionality, which is named Hughes Phenomenon. To deal with this problem, many non-linear methods have been proposed such as kernel support vector machines, sparse multinomial logistic regression, neural networks and so forth. Although these methods can achieve better accuracy than traditional linear mothods, there are still limitations: (1) they mainly make use of the spectral information of the pixel, regardless of the spatial context of the pixel. (2) Directly using the low level spectral bands as the feature is not discriminative enough for classification.

Considering these limitations described above, deep learning methods are promising to alleviate them. Deep neural networks can be viewed as a strong non-linear function, which has advantages over traditional methods for handling the HSI data with high dimensionality. Another superiority of deep convolution networks is that the receptive fields of CNN models take the neighboring pixels into consideration inherently. The spatial feature provides context information and is an important clue for the hyperspectral image pixel-wise classification.

Although deep convolutional networks take spatial context into account, traditional CNN models merely exploit the fixed convolution kernel size. However, the hyperspectral image land cover class distribution is complicated, traditional CNN with fixed kernel size is not flexible enough. Convolution with different spatial context size may capture more discriminative context feature for HSI pixel classification.

In addition to these issues above, for hyperspectral image classification task, merely less than 10% of the samples can be used for training. However, deep learning models usually have huge amounts of training parameters. So if the training samples size is too small, deep models tend to be overfitting. Recently proposed deep network architectures like resnet [1], and densenet[2] are very deep, even beyond one hundred layers. These networks are effective for huge datasets such as Imagenet and coco. However, these networks are too complex for HSI images. The training samples are insufficient, so these models suffer from overfitting problem.

To address the problems depicted above, in this paper, we

design a deep attention inception CNN architecture named AI-NET. The contributions of this work can be summarized as follows:

(1) We introduce a novel deep attention inception module. This module can adaptively focus on different spatial context by using different convolution kernel size. With respect to different datasets with unique characteristics, the module can learn special attention pattern to improve the classification performance.

(2) We design an effective two layer attention inception module based deep network architecture. We address the overfitting problem by designing samll scale (shallow) but large capacity (multiple convolution kernels with attention) CNN architecture.

2. RELATED WORK

Recent machine learning methods devised for hyperspectral image classification can be divided into two categories: traditional classification methods and deep learning based methods.

In recent years, some advanced classifier such as local Fisher discriminant analysis [3], and kernel support vector machines[4]. SVM is a kind of maximum margin based classifier, and has been widely applied to HSI classification task. By using the kernel function, kernel SVM can classify the HSI data in higher dimensionality space. Some improved Neural networks (NN) classifiers such as applied radial basis function NN for HSI classifiation. However the limitation of these methods is that the low level feature is not representative enough for complicated HSI images.

Recently, great progress has been made in HSI classification owing to the deep learning methods, especially the deep convolutional neural networks. Deep CNN is effective for high level feature extraction and has shown its power in many computer vision tasks like [5] and [6]. CNN is composed of convolution, activation and pooling layers. By stacking these layers, CNN can learn hierarchical representation of the images. Deep learning based methods, such as stacked autoencoders [7] and supervised deep CNN [8] are used to extract the spectral features of HSI, and obtained promising performance. Principal component analysis (PCA) is exploited as the preprocess of the deep CNN in work [9].

3. OUR METHOD

The whole network is illustrated by Fig. 1. The main process of the system is as follows. In the training stage, we extract 7×7 image patches at every HSI pixel as the input images. Then the image patches are fed into a convolution layer with 3×3 kernel size and a max pooling layer. Next, two attention inception modules with residual connection are exploited to learn the dynamic spatial context features. All convolution layers are activated by Relu layer. Finally, we use fully connected layer and softmax layer as the classifier for predicting pixel-wise class. The proposed framework can be easily trained in an end-to-end manner.

3.1. Multiple kernel Inception Module

Inception module is proposed to learn the optimal sparse structure representation. As illustrated by Fig. 1, the inception mudule is composed of three branch convolution operation with three different branches: (1). one 1×1 kernel; (2). one 3×3 kernel; (3). two 3×3 kernels. Among them, the branch with two 3×3 kernels is the approximation of one 5×5 kernel with less parameters and computation cost. If we denote the k-th feature map as $\mathrm{h^k}$, the three branch convolution outputs are $h^k_{1\times1}$, $h^k_{3\times3}$ and $h^k_{5\times5}$ respectively. The filters $w_{1\times1}$, $w_{3\times3}$ are corresponding to 1×1, 3×3 convolution kernel respectively. $*$ denotes the convolution operation, and b^k_n represents the bias corresponding to the weights. Then the operation of the multiple kernel convolution can be formulated as follows:

$$h^k_{1\times1} = relu(w_{1\times1} * x + b^k_{1\times1}), \tag{1}$$

$$h^k_{3\times3} = relu(w_{3\times3} * x + b^k_{3\times3}), \tag{2}$$

$$h^k_{5\times5} = relu(w_{3\times3} * (relu(w_{3\times3} * x + b^k_{3\times3})) + b^k_{3\times3}). \tag{3}$$

The Multi-kernel convolution network takes advantage of multiple spatial context to extract features with different characteristics. The three branch outputs $h^k_{1\times1}$, $h^k_{3\times3}$ and $h^k_{5\times5}$ are the inputs of the next attention mechanism.

3.2. Attention Inception Model

The attention inception model dynamically focus on the multiple spatial kernel convolution outputs by attention vectors. Traditional inception module concatenates different branch outputs together as the next layer inputs. In our proposed method, we use attention mechanism to adaptively encode different spatial size information to learn more representative features. We denote the attention vector as $[w_1, w_2, w_3]$, and the output of attention inception model of the k-th feature map is defined as h^k_a

$$h^k_a = w_1 * h^k_{1\times1} + w_2 * h^k_{3\times3} + w_3 * h^k_{5\times5}. \tag{4}$$

Then the attended feature maps are fed into the next attention inception module, and the residual connection is used to make the loss back propagation process more effectively.

3.3. AI-NET Architecture

The first attention inception module output h^k_a is taken as the input of the second attention inception module. After the

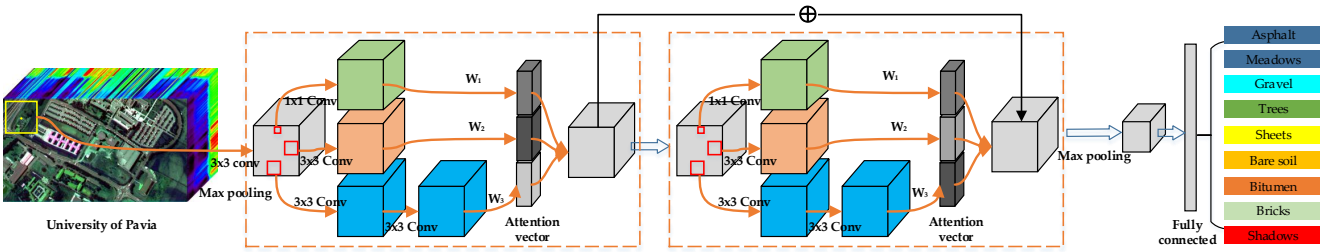

Fig. 1. The whole detection framework.

Table 1. AI-NET NETWORK ARCHITECTURE

Type	Kernel size	Feature maps
Convolution	3x3	64
Max pool	2x2	64
Inception1/Convolution1	1x1	128
Inception1/Convolution2	3x3	128
Inception1/Convolution3_1	3x3	128
Inception1/Convolution3_2	3x3	128
Inception2/Convolution1	1x1	128
Inception2/Convolution2	3x3	128
Inception2/Convolution3_1	3x3	128
Inception2/Convolution3_2	3x3	128
Max pool	2x2	128
Fully connected	-	-

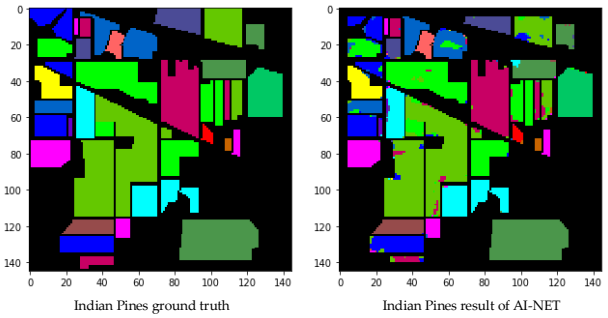

Fig. 2. The classification map of our method and the ground truth label map.

same process with the first module, the output of the second attention inception module is denoted as $h2_a^k$. Then the residual connection is exploited as follows:

$$h3_a^k = h_a^k + h2_a^k, \tag{5}$$

Then $h3_a^k$ is fed to a fully connected layer and the prediction class is output by softmax layer. The full architecture is shown in table 1.

4. EXPERIMENTS

To evaluate the proposed AI-NET, we trained and tested it on two public HSI classification datasets: the Indian Pines dataset and Salinas dataset. There are 16 land cover classes in Indian Pines dataset, and 16 land cover classer of urban area in University of Pavia dataset. For conveniently comparing our method with other published methods, in the IN dataset, we randomly select 200 samples of each annotated class for training. For the UP dataset, we also obrtain 200 randomly chosen annotated data for training. We set the patch size to 7×7 in both dataset. Stochastic Gradient Descent (SGD) is used to train the model with the following parameters: The learning rate is set to 0.001, and the weight decay is 0.0004. The batch size we use in our experiment is 100. Batch normalization is not used in our designed network.

We compare our method with traditional advanced machine learning methods like radial basis function kernel SVM (F-SVM), extended morphological profiles SVM (EMP-SVM) [10], EMP-CNN [11] and Gabor-CNN [12]. All experiments are run 10 times with the same training parameters and random initial weights. Overall accuracy (OA) and Kappa are used as the evaluation measurements for all the compared methods.

Table 2. TESTING CLASSIFICATION RESULTS WITH 200 TRAINING SAMPLES

	Salinas dataset		IN dataset	
Methods	OA(%)	Kappa	OA(%)	Kappa
RBF-SVM	83.09	81.07	58.01	52.07
EMP-SVM [10]	85.90	84.02	69.34	64.56
EMP-CNN [11]	87.04	85.43	86.48	84.23
Gabor-CNN [12]	92.02	91.07	89.02	86.07
AI-NET(ours)	**94.64**	**92.73**	**93.07**	**91.75**

The experimental results of the IN dataset are shown in Table. 2. From the result we can see that our designed network with attention inception module is effective for HSI data classification. From the results, we can see that traditional methods like the modified SVM achieves poor performance with only 200 samples. CNN based methods are more effective for its more dicriminative feature learning power, and improve the accuracy by almost 20%. The proposed method achieves the best classification performance, which indicates the active effect of the proposed multi-spatial context model.

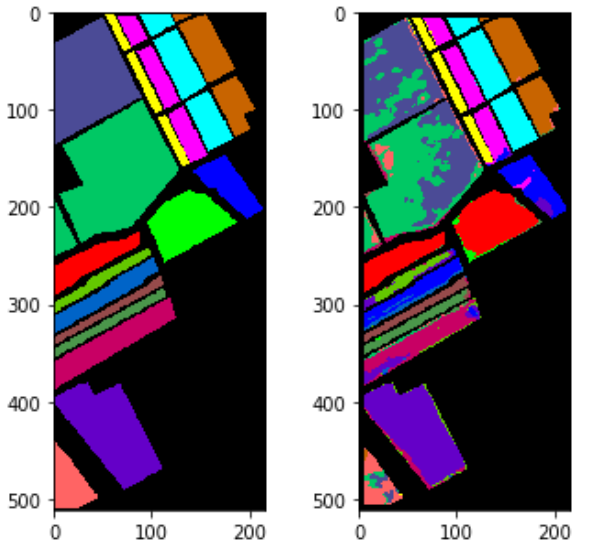

Fig. 3. The classification map of our method and the ground truth label map.

The qualitative land cover classification result on IN dataset is shown in Fig.2.

From Table. 2, we can see the comparison results of the Salinas dataset. From the result we can see that traditional modified methods like SVM can achieve about 84% accuracy. The land cover class ditribution of Salinas dataset is more pure and simple, so traditional mathods as well as the CNN based methods can obtain better performance than the IN dataset. In this dataset, our AI-NET also achieves best accuracy, which shows that the adaptive attention inception network is robust to different land cover class distribution. The qualitative land cover classification result on Salinas dataset is shown in Fig.3.

5. ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61773316 and 61379094, and the Open Research Fund of Key Laboratory of Spectral Imaging TechnologyChinese Academy of Sciences.

6. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[3] Wei Li, Saurabh Prasad, James E Fowler, and Lori Mann Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1185–1198, 2012.

[4] Farid Melgani and Lorenzo Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[5] Qi Wang, Jia Wan, and Yuan Yuan, "Deep metric learning for crowdedness regression," *IEEE Transactions on Circuits and Systems for Video Technology*, doi:10.1109/TCSVT.2017.2703920, 2018.

[6] Qi Wang, Junyu Gao, and Yuan Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 230–241, 2018.

[7] 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, 2014.

[8] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015.

[9] Jun Yue, Wenzhi Zhao, Shanjun Mao, and Hui Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.

[10] Mathieu Fauvel, Jón Atli Benediktsson, Jocelyn Chanussot, and Johannes R Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

[11] Erchan Aptoula, Murat Can Ozdemir, and Berrin Yanikoglu, "Deep learning with attribute profiles for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1970–1974, 2016.

[12] Yushi Chen, Lin Zhu, Pedram Ghamisi, Xiuping Jia, Guoyu Li, and Liang Tang, "Hyperspectral images classification with gabor filtering and convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2355–2359, 2017.