Journal of Scientific & Industrial Research
Vol. 79, December 2020, pp. 1087-1094

Scene based Classification of Aerial Images using Convolution Neural Networks

Palak Mahajan[1]*, Pawanesh Abrol[1] and Parveen K Lehana[2]

[1]Department of Computer Science & IT, [2]Department of Electronics, University of Jammu, J&K, India

Received 12 June 2020; revised 06 July 2020; accepted 15 October 2020

The advent of computer vision and evolution of high-end computing in remote sensing images have embellish various researchers for unprecedented development in remotely sensed aerial images. The requirement to extract essential information stimulated anatomization of aerial images for its usefulness. Deep learning provides state of the art solutions for widely explored visual recognition system and has emerged as an evolutionary area, being applicable to large scale image processing applications. Convolutional Neural Networks (CNNs), an essential component of deep learning algorithms consists of increasing the depth and connections in the processing layers to learn various features of data at different abstract levels. In this paper, we present an outlook for classifying and extracting the features of aerial images using CNN. We propose a CNN architecture based on various parameters and layers for classification. CNN has been evaluated on two publicly available aerial data sets: UC Merced Land Use and RSSCN7. Experimental results show that the proposed CNN architecture is competent and efficient in terms of accuracy as performance evaluation parameter in comparison with conventional classifiers like Bag of Visual Words (BOVW).

Keywords: CNN, Deep learning, Feature extraction, Image classification

Introduction

The fast track evolution in remote sensing led to immense development in fostering techniques to precipitate remotely sensed aerial images with high resolution. Publically availability of satellite aerial images for research purpose instigated wide-ranging application areas including urban planning, terrain monitoring[1], flood, surveillance[2], agriculture[3], etc. Thereby plugging demand for explicit analysis on aerial images to extract essential information for designing structured semantics tasks[4] such as Scene based classification. With an insight to label scenes into Land use and land cover (LULC) classes[5], scene based classification categorizes the class labels according to the representation of the land type such as highway, fields, etc. Aerial images require high resolution data, demanding immense analysis to provide accurate classification. Spatial features are pre-eminent for learning and extraction in high resolution aerial images. The challenge being extracting these perplexing spatial feature representations from high resolution aerial images.[6] Such classification requires require an assembly of essential core representations of the features in an intelligent machine learning system. The expansion of deep learning in the computer vision has been a breakthrough in machine learning.

The demand of computer vision system is to yield essential internal representations through feature extractor known as features, the output of which is fed to a trainable classifier. These extracted features direct the system to solve classification problem into categories, independently of scale, illumination, position, or clutter. The exhaustive search to learn features has led to the growth of deep learning. Deep learning can be defined as a neural net consisting of several hidden layers and in comprehensive term, as a learning model facilitated with layered feature extraction.[7] Deep learning architectures initialize with original feature as its input. It transforms these features into abstract features by extracting layer by layer traversing deeply into the multi-layered network, thereby, enhancing the performance of the net. In other words, these deep architectures are capable of extracting features and shaping them into abstract representations. Support Vector Machine (SVM)[8], Bayesian networks[9], Bag of Visual Words (BOVW)[5], K-Nearest Neighbours (KNN) and many learning algorithms resolve the classification problem. However, the limitations include restriction on designing and developing feature extractor with poor generalization.

*Author for Correspondence
E-mail: palak.mahajan18@gmail.com

An image comprises of millions of pixels. To understand an image, one needs to comprehend how these pixels are clubbed together to create an image. Generally, the pixels are integrated into edglets, edglets into motifs, motifs into regions, regions into objects, and objects into scenes.[10] Hence, a computer vision system requires having numerous trainable stages, one for each level representation in the feature hierarchy to perform an optimum classification. Convolutional Neural Networks (CNNs)[11–13] came into counting while describing trainable multi-layer deep learning architectures. The input and output at each layer are defined as feature maps. Greater the number of multiple stages in a CNN, higher is the learning rate of the multi-level feature hierarchies. Its architecture provides distinctive characteristics that enable CNN to do minimal pre-processing task for performing operations like segmentation, feature extraction, classification, object recognition, etc. Unlike conventional pattern recognition approaches in which prior problem domain knowledge is required to select the precise algorithm, a simple knowledge of problem domain is sufficient to extract the features in case of CNN. Images are defined using local and global features such as textures, colour, illumination, etc. Implementation of CNNs enables the processing of adequate amount of training and classification required for aerial data. This minimizes the time, cost and computational consumption. CNNs have an extensive property of learning features from an unambiguous specified dataset with one modality and applying it other dataset of different modality with some additional training. Signifying a pre-trained CNN works better than a net trained from the scratch.[14]

In this paper, we will explore the prospect of learning and classifying explicitly in deep neural networks i.e. CNN for classification of aerial images. CNN in combination with various layers such as normalization[15], pooling, dropout, softmax, etc. have been implemented. The net being designed has limited number of layers, and each layer is designed with building blocks such as convolution, normalization, pooling, etc. Experiments are conducted on two benchmark aerial datasets with widely different characteristics i.e. UC Merced Land Use[5] and RSSCN7[16] datasets containing 2100 and 2800 images respectively. The potential of the CNN is evaluated based on classification accuracy.

The remainder of this paper is structured as follows: Section 2 reviews the literature. Section 3 presents the detailed theory of the network model structure of CNN. In Section 4, the experiments and results have been discussed. Conclusion is presented in Section 5.

Related Work

Since the beginning of current decade there has been intense research on image classification through deep networks. The vast architecture of deep networks like CNNs made the researchers to provide emphasis on the use of suitable learning models. Deep learning techniques infused with SVM and regression framework were proposed.[7] The framework integrated worked on joint spectral–spatial classification. Experiments showed an improved accuracy compared with other techniques like PCA. CaffeNet and GoogLeNet architectures of CNNs were implementation for aerial image classification.[17] The network worked on minimizing design time and overfitting problems. The results summarized improved classification accuracy with additional training of networks with pre- trained larger dataset. CNN has been implemented for learning specific spatial features from aerial images with an objective of determining corresponding hierarchical structures in images.[18] The layers were arranged systematically so that the initial five layers extracted the visual features while the last layer was accountable for classification purpose. An object-based classification on a convolution neural network (CNN) has been proposed.[13] The classification was implemented in two-stages, GoogleNet architecture and NVidia Digits were used for training. The results indicated accurate identification to locate the regions in images which correspond to the categories on which the CNN was trained. However, the net can be enhanced by employing multi-GPU configuration to reduce computation time. ImageNet[19] database was utilized to determine the effect of CNNs depth on its parameters like accuracy in the large-scale image recognition.[20] In this work, the depth of the CNN was stretched to around 16–19 weight layers that provided satisfactory classification accuracy. Another deep learning-based classification was introduced based on vehicle detection and counting in aerial images while employing CNNs for regression of vehicle spatial density map across the aerial image.[21] Experimental work done on Munich and Overhead Imagery Research datasets provided higher precision and recall rates. Another deep learning technique were used for feature extraction for scene classification and adapted feature selection as feature reconstruction.[16] The

reconstructed features were further used as discriminative features for image representation. DBN enabled the reduction of reconstruction error and experiments validated the performance of the model. CNN framework was applied for 2-D cardiac magnetic resonance (MR) images from under sampled data for image reconstruction in real time application.[22] Deep neural networks framework known as DeepID3 was proposed for face recognition.[2] The framework was based on VGGNet[3] and GoogLeNet[23] architectures. Genetic algorithm[24] was designing with deep neural networks.[25] The exponential increase in CNN layers were optimized using genetic algorithm to efficiently traverse the large search space. Each component of individual layers was defined through training the net followed by evaluating its validation set that computed its recognition accuracy.

An unsupervised learning with multi-stage hierarchies was proposed that worked on sparse convolutional features.[26] The model controlled the redundancy between feature vectors at neighbouring locations by providing highly di-verse filters through convolutional training. An improved efficiency of the over-all visual representation was achieved. A greedy layer-wise unsupervised frame-work coupled with single-layer deep CNN was implemented.[27] The frame-work engrained on sparse representations of multi and hyperspectral imagery. The results illustrated dynamic performance of the framework in very high-resolution images with diverse classification scenarios. An aerial scene classification using dense low-level feature sets has been described.[28] Unsupervised learning on unlabelled features sets operates on encoding, pooling and extraction to generate image representations.

In brief we can approximate that there is scope of exploring deep learning models for feature extraction and classification based on the image under study in many aspects. As observed different architectures have been proposed with varying accuracy parameters and different classification techniques. The research involved in this paper explores CNNs in depth with diverse architecture and number of layers to achieve better classification performance.

Proposed Classification Framework

After going through the extensively detailed literature and understanding the different CNN architectures proposed and implemented by different researchers for a variety of applications, we propose a classification framework comprising of various CNN layers shown by Fig. 1. It has been observed that the variation in the order of layers and careful selection of parameters as per system configuration and output requirement impacts the complexity and performance of the classification system.

Here we propose a CNN based classification framework comprising of six layers: the first three are convolution layers and followed by fully-connected and softmax layers. The output of softmax layer is fed to classification layer which produces a distribution over multiple class labels as defined. The framework has been experimentally analysed using UC Merced Land Use and RSSCN7 datasets. The concept of multinomial logistic regression is extended by the proposed CNN which maximizes the average across training cases of the log-probability of the correct label under the prediction distribution.[29] It is achieved by using mini-batch gradient descent with momentum. The kernels[29] of all three convolutional layers are interconnected with all other kernel maps in

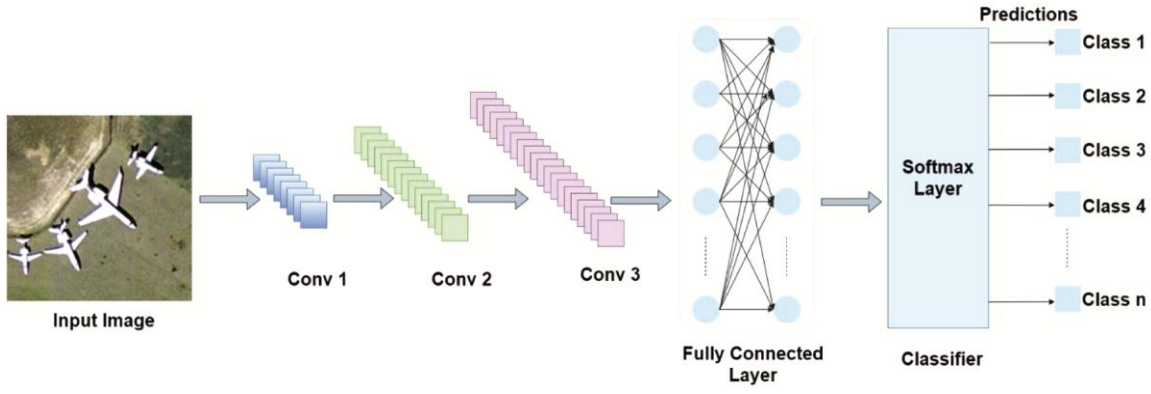

Fig. 1 — Proposed CNN architecture with six layers: the initial three are convolutional, followed by fully-connected and softmax; The output of the softmax layer is fed to classifier layer that produces the probability distribution over the possible class labels

the subsequent layer. The nodes in the fully-connected layers are connected to all nodes in the previous layer. Batch normalization layers follow the first and second convolutional layers. Max-pooling[11,31] layers follow both normalization layers and the third convolutional layer. The ReLU[32] non-linearity layer is applied within all the layers to blend non-linearity into the network, the mathematical form of which has been defined in Eq. 3. To begin with, the first convolution layer filters the input image, which can be of varied size depending on the dataset being used.

The feature map generated by learning through the first and second convolution layer while training the network is illustrated by Fig. 2. It comprises of strongest activation channels with each kernel representing feature specifications. The features of various individual class labels are defined and identified by extracting layer by layer each convolutional kernel. The second convolutional layer includes max-pooling and batch-normalized operations that takes the output of the first convolutional layer as input and filters it kernels. The third convolutional layer includes max-pooling and batch-normalized operations that have kernels connected to the outputs of the former convolutional layer. Subsequently, the fully-connected layer has all nodes connected to each other from the previous convolution layer. The output of fully-connected layer is then distributed to the softmax layer on which dropout has been applied that consists of setting to zero the output of each hidden node with probability 0.5.The dropout layer is added to remove the irrelevant nodes in the net thereby the reducing the overall size of CNN. Finally, the network has the classification layer linked with softmax layer that provides the probability distribution over the possible classes.

Preliminary investigations were performed in which the CNN model was run multiple times giving

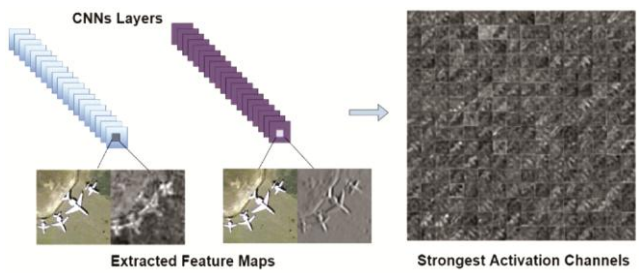

Fig. 2 — A visualization of feature maps learned by the convolution layer in our proposed architecture.

different range of parameters. Based on different range of values, the proposed model is trained with generated optimum values thereby reducing training error. The training is done using stochastic gradient descent with a batch size of 256 examples, momentum of 0.9, and weight decay of 0.0002. The weight decay is applied to reduce training error. Also, the training was regularized by weight decay (the L2 penalty multiplier set to 10−4) and dropout regularization for the fully-connected and softmax layers (dropout ratio set to 0.5). The learning rate is initially set to 1.0 e−4, and then decreased by a factor of 10 when the validation set accuracy stopped improving. Stochastic gradient descent has been used as learning function during the network training with 50 epochs respectively. Our objective is to achieve high precision rate for all the classes so that the network can be used for aerial image classification. To obtain the fixed-size CNN input images, they are randomly cropped from rescaled training images. The deployment of our proposed CNN architecture for classification of aerial images for datasets UCMerced Land Use and RSSCN7 respectively are listed in Table 1. The configuration details of various layers used in the proposed CNN has been described in Table 1. The layers comprise of 3 × 3 convolutions followed by rectified linear unit (ReLU), max-pooling and batch normalization respectively. A 50% dropout ratio is applied along with softmax and classification layer. The size of the output of classification layer is set to 21 and 7 depending upon the number of classes for the two varied datasets.

Table 1 — Detailed architecture of the proposed CNN

Layers		Configuration
Layer 1	Conv	[5 × 5 × 64]
	ReLU	Max (0, x)
Layer 2	Conv	[3 × 3 × 128]
	Max Pooling	[2 × 2]
	ReLU	Max (0, x)
Layer 3	Conv	[3 × 3 × 256]
	Max Pooling	[2 x 2]
	Batch Normalization	Alpha = 0.01, beta = 0.5
	ReLU	Max (0, x)
Layer 4	Fully Connected	Weight Size: [1158, 256]
	Max Pooling	[2 × 2]
	ReLU	Max (0, x)
Layer 5	Softmax	Weight Size: [256, 128]
	Dropout	Activation: 50%
	ReLU	Max(0, x)
Layer 6	Classification	Weight Size: [128, 21] / [128, 7]
	Dropout	Activation: 50%
	ReLU	Max (0, x)

Results and Discussion

On the basis of the experiment being conducted, the output has been generalized. The significant issues involved with wide diversity of input data dimensionality, number of classes, and amount of available labelled data for feature extraction in land-use classification for aerial images are being studied. The prime focus of the experiment is to address the relevant issues such as impact of number of classes, depth of CNN layers and the learned hierarchical representations while training the network. The experiment is setup on Intel(R) Core (TM) i7-4770 CPU @3.40 GHz processor with 6 GB RAM. In each experiment we have selected randomly 70% images per category for training and rest for testing. The platform used to implement CNNs is MATLAB R2018a on Windows 10.

Data Collection

We validate the aerial scene classification on two datasets i.e. UCMerced Land Use and RSSCN7 dataset respectively. In UCMerced, the data set comprises of manually extracted images from the USGS National Map Urban Area Imagery collection. It comprises of twenty-one aerial scene categories with 256 × 256 colour images with 1-ft/pixel resolution. The images in this dataset cover overlapping categories and each category contains 100 images thereby creating a dataset of 2100 images. Few ground truth images for each twenty-one land-use categories are represented by Fig. 3. The extensive categories of the dataset are defined as follows: agriculture, airplane, baseball diamond, beach, building, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court respectively.

The aerial data set RSSCN7 contains 2800 remote sensing aerial images that comprises of seven categories, i.e., grass, forest, farm, parking lot, residential region, industrial region, and river and lake. There are around 400 images collected in each category via Google Earth. Each image consists of 400 × 400 pixels. This data set is reasonably challenging due to the wide diversity of overlapping scene images that are captured under varying weather conditions. Some ground truth RSSCN7 images for seven categories are represented by Fig. 4.

Observation

UCMerced Land Use and RSSCN7 datasets are adopted to test the performance of the system in terms of classification accuracy. Here we have trained the proposed CNN for 30 epochs and calculated the accuracy and loss profile of the classification system for both datasets, shown by Fig. 5.While training the proposed CNN, it has been observed that there is a smooth increasing accuracy i.e. with the increase in

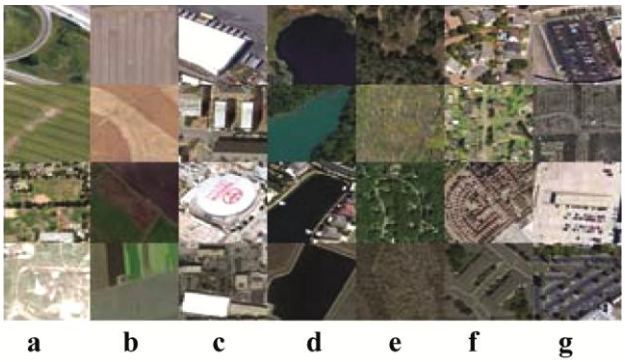

Fig. 4 — Sample images from the RSSCN7 ground truth data set containing 400 images for each 7 categories, from which four samples per category has been shown. The following are the labeled category: **a** Grass, **b** Field, **c** Industry, **d** Riverlake, **e** Forest, **f** Residential, **g** Parking, respectively

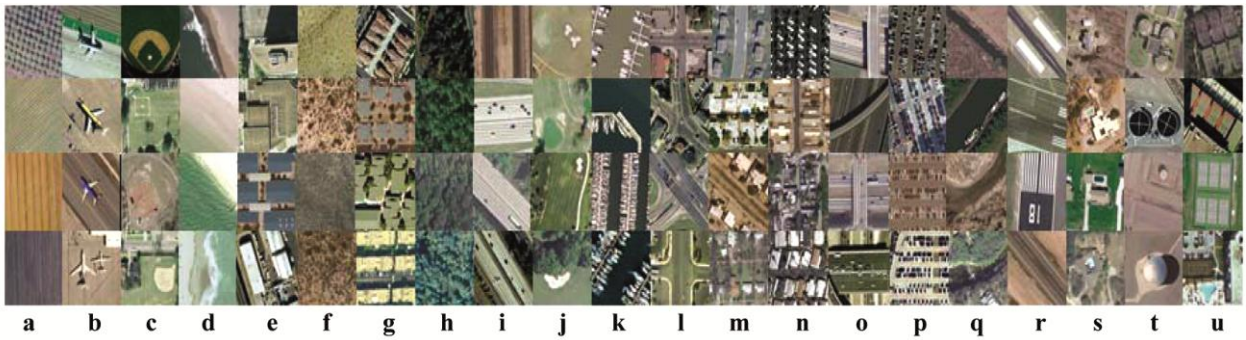

Fig. 3 — UCMerced ground truth data set contains 100 images for each 21 category, from which four samples per category has been shown. The following are the labelled category: **a** Agriculture, **b** Airplane, **c** Baseballdiamond, **d** Beach, **e** Building, **f** Chaparral, **g** Dense residential, **h** Forest, **i** Freeway, **j** Golfcourse, **k** Harbor, **l** Intersection, **m** Medium residential, **n** Mobile homepark, **o** Overpass, **p** Parkinglot, **q** River, **r** Runway, **s** Sparse residential, **t** Storagetanks, **u** Tenniscourt, respectively

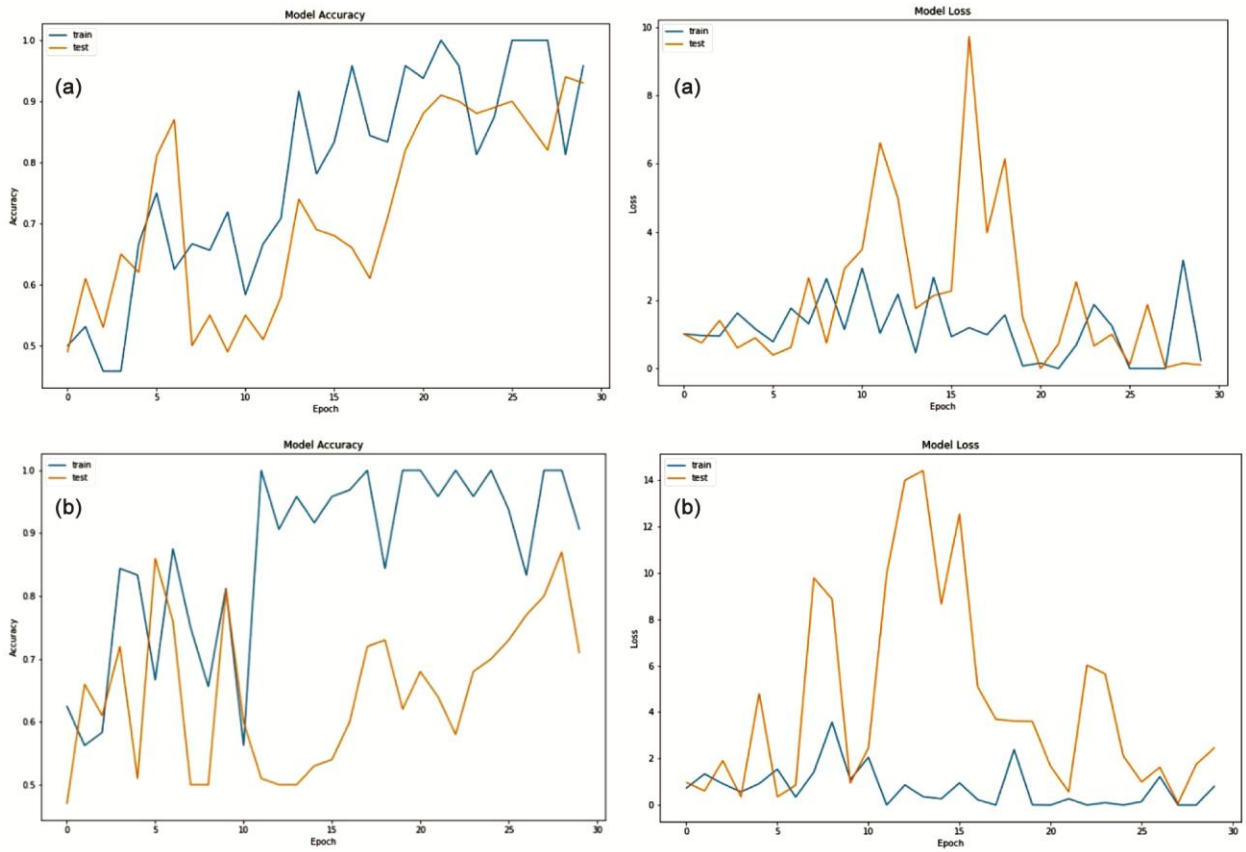

Fig. 5 — Graphical profiles for accuracy and loss during training the proposed CNN for (**a)** UCMerced Land Use and (**b)** RSSCN7 datasets respectively

number of epochs, the accuracy increases. The optimum performance of CNN for has been achieved at 29th epoch with an accuracy of 92.65% with loss of 0.1589 for UCMerced Land Use and 87% accuracy with loss 1.7503 for RSSCN7 as shown in the graphical profile of accuracy and loss respectively.

Also, experiments for performance parameters over classical classification techniques like Latent Dirichlet Allocation (LDA)[33], Vector of Locally Aggregated Descriptors (VLAD)[34], Spatial Pyramid Matching (SPM)[35], BOVW, Improved Fisher kernel (IFK)[36] for both datasets are computed. The performance of proposed CNN architecture in terms of accuracy with other classical classification techniques is concluded by Table 2. According to the table, training the datasets with proposed CNN architecture accuracy of 76.60% and 74.56% is achieved which have vividly increased the performance of the classification framework. The accuracy rates for LDA, VLAD, SPM, IFK and BOVW as presented in Table 2 range from 61% – 77%. As observed, the CNN architecture showed high performance with increased accuracy compared to classical classification techniques. We can observe that the performance of UCMerced Land Use dataset is slightly higher than the RSSCN7 dataset reason being the former dataset have a greater number of categories than the later one. Hence, it can be inferenced from the experimental results that greater the number of categories and deeper the number of layers is, higher is the net performance.

Table 2 — Experimental results of CNN architectures over UCMerced Land Use and RSSCN7 testing datasets

S.No.	Classifier	UCMerced Land Use (%)	RSSCN7 (%)
1.	LDA	64.72	71.36
2.	VLAD	75.38	74.42
3.	BOVW	78.25	76.43
4.	SPM	61.63	64.25
5.	IFK	79.81	77.82
6.	Proposed CNN	92.65	87.00

However, increasing width of layers caused overhead as it takes more time for processing due to which the size of CNN is limited. Also, with more classes in UCMerced Land Use dataset, there is quadratic time hiked computation time for RSSCN7

dataset. Hence, due to high computational time of CNN with run-time being major concern in CNN applicability, few layers have been attached in the proposed CNN architecture.

Conclusions

This paper exploits the structure of deep convolution networks. We have proposed a CNN architecture comprising of various layers and have implemented and trained it from the scratch. The performance evaluation is done on two challenging datasets UCMerced Land Use and RSSCN7. The experimental results validated utility over classical classification techniques. It has been observed that adding layers into the CNN improves the classification accuracy substantially. Further, the deeper the network is, higher is the classification performance. However, limited system resources like computing power and time has limited our net in terms of how far it could have gone for training each net. Due to which we have implemented the net for few epochs for each net before it plateaued. In the future work, the experimental analysis can be extended using graphics processing units (GPU) to accelerate the feature learning process. Further this approach shall be extended to perform texture-based object detection with high-level spatial information as part of the feature extraction process.

References

1 Mou L, Zhu X, IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network, *arXiv preprint* (2018) arXiv:1802.10249.

2 Sun Y, Liang D, Wang X, Tang X, DeepID3: face recognition with very deep neural networks, *arXiv preprint* (2015) arXiv:1502.00873.

3 Samaniego L, Schulz K, Supervised classification of agricultural land cover using a modified k-NN technique and landsat remote sensing imagery, *J Remote Sens*, **1** (2009) 875–895.

4 Huang L, Liu B, Li B, Guo W, Yu W, Zhang Z, Yu W, A dataset dedicated to sentinel-1 ship interpretation, *IEEE J Sel Top Appl Earth Obs Remote Sens*, **11(1)** (2018) 195–208.

5 Yang Y, Newsam S, Bag-of-visual-words and spatial extensions for land-use classification, *Proc of SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010, 270–279.

6 Mahajan P, Abrol P, Lehana P K, Effect of blurring on identification of aerial images using convolution neural networks, *Lect Notes Electr Eng*, **597** (2020) 469–484.

7 Arel I, Rose D C, Karnowski T P, Deep machine learning - a new frontier in artificial intelligence research, *IEEE Comput Intell Mag* **5(4)** (2010) 13–18.

8 Chen Y, Lin Z, Zhao X, Wang G, Gu Y, Deep learning-based classification of hyperspectral data, *IEEE J Sel Top Appl Earth Obs Remote Sens* **7(6)** (2014) 2094–2107.

9 Ghamisi P, Plaza J, Chen Y, Li J, Plaza A, Advanced spectral classifiers for hyperspectral images, *IEEE Geosci and Remote Sens Mag* **5(1)** (2017) 8–32.

10 Bishop C M, Pattern Recognition and Machine Learning, Springer, 2007, New York.

11 Bengio Y, Learning deep architectures for AI, *Found Trends Mach Learn* **2(1)** (2009) 1–127.

12 Bouvrie J, Notes on convolutional neural networks, 2006.

13 Sevo I, Avramovic A, Convolutional neural network based automatic object detection on aerial images, *IEEE Geosci and Remote Sens Lett* **13(5)** (2016) 740 – 744.

14 Kobayashi F K, Mattos A B, Gemignani B H, Macedo M G, Experimental analysis of citrus tree classification from UAV images, *IEEE International Symposium on Multimedia* 2019.

15 Ioffe S, Szegedy C, Batch normalization: accelerating deep network training by reducing internal covariate shift, *arXiv preprint* (2015) arXiv:1502.03167.

16 Zou Q, Ni L, Zhang T, Wang Q, Deep learning based feature selection for remote sensing scene classification, *IEEE Geosci Remote Sens Lett*, **12(11)** (2015) 2321–2325.

17 Castelluccio M, Poggi G, Sansone C, Verdoliva L, Land use classification in remote sensing images by convolutional neural networks improving spatial, *arXiv preprint* (2015) arXiv:1508.00092v1.

18 Nogueira K, Miranda W O, Santos J A, Improving spatial feature representation from aerial scenes by using convolutional networks, *Proc SIBGRAPI Conference on Graphics, Patterns and Images*, 2015, 289–296.

19 Krizhevsky, Sutskever I, Hinton G E, ImageNet classification with deep convolutional neural networks, *Adv Neur Inf Proces Syst*, (2012) 1106–1114.

20 Simonyan K, Zisserman A, Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2015) arXiv:1409.1556.

21 Tayara H, Soo KG, Chong KT, Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network, *IEEE Access*, **6** (2017) 2220–2210.

22 Schlemper J, Caballero J, Hajnal J V, Price A N, Rueckert D, A deep cascade of convolutional neural networks for dynamic mr image reconstruction, *IEEE Trans Med Imaging*, **37(2)** (2018) 491–503.

23 Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A, Going deeper with convolutions, *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 1–9.

24 Sambyal P, Abrol P, Lehana P, Optimization of light switching pattern on large scale using genetic algorithm, Int J Sci Tech Adv, **3(1)**, (2017) 19–23.

25 Xie L, Yuille A, Genetic CNN.*arXiv preprint* (2017) arXiv:1703.01513.

26 Kavukcuoglu K, Sermanet P, Boureau Y L, Gregor K, Mathieu M, LeCun Y, Learning convolutional feature hierarchies for visual recognition unsupervised deep feature, *Proc ACM Conference on Advances in Neural Information Processing Systems*, 2010, 1090–1098.

27 Romero A, Gatta C, Valls G C, Unsupervised deep feature extraction for remote sensing image classification, *IEEE Trans Geosci Remote Sens*, **54(3)** (2016) 1349–1362.
28 Cheriyadat A M, Unsupervised feature learning for aerial scene classification, *IEEE Trans on Geosci Remote Sens*, **52(1)** (2014) 429–451.
29 Cao F, Yang Z, Ren J, Ling WK, Extreme sparse multinomial logistic regression: a fast and robust framework for hyperspectral image classification, *arXiv preprint* (2017) arXiv:1709.02517.
30 Aydogdu M F, Celik V, Demirci M F, Comparison of three different cnn architectures for age classification, *Proc IEEE International Conference on Semantic Computing*, 2017, 372–377.
31 Nagi J, Ducatelle F, Di Caro G A, Ciresan D, Meier U, Giusti A, Nagi F, Schmidhuber J, Gambardella L M, Maxpooling convolutional neural networks for vision based hand gesture recognition, *In Proc IEEE International Conference on Signal and Image Processing Applications*, 2011, 343–349.
32 LeCun Y, Kavukcuoglu K, Farabet C, Convolutional networks and applications in vision, *Proc IEEE International Symposium on Circuits and Systems*,2010, 253–256.
33 Blei D M, Ng A Y, Jordan M I, Latent dirichlet allocation, *J Machine Learn Res*, **3** (2003) 993–1022.
34 Negrel R, Picard D, Gosselin P H, Evaluation of second-order visual features for land-use classification, in *International Workshop on Content-Based Multimedia Indexing*, 2014, 1–5.
35 Lazebnik S, Schmid C, Ponce J, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *Proc IEEE Conference on Computer Vision and Pattern Recognition*, **2**, 2006, 2169–2178.
36 Perronnin F, Sanchez J, Mensink T, Improving the fisher kernel for large-scale image classication, *Proc European Conference on Computer Vision*, 2010, 143–156.