An Efficient Convolutional Neural Network for Remote-Sensing Scene Image Classification

Muhammad Ashad Baloch[1*], Sajid Ali[2], Mubashir H.Malik[3], Aamir Hussain[4], Abdul Mustaan Madni[1]

[1] National College of Business Administration & Economics Multan, Pakistan.
[2]Department of Information Science, University of Education Lahore, Pakistan.
[3] Institute of Southern Punjab Multan, Pakistan.
[4] Muhammad Nawaz Shareef University of Agriculture Multan, Pakistan.

* Corresponding author. Tel.: +923056976008; email: Ashad5765@gmail.com
Manuscript submitted December 12, 2019; accepted January 13, 2020.
doi: 10.17706/jcp.15.2.48-58

Abstract: Deep neural networks are providing a powerful solution for remote-sensing scene image classification. However, a limited number of training samples, inter-class similarity among scene categories, and to get the benefits of multi-layer features remains a significant challenge in the remote sensing domain. Many efforts have been proposed to deal the above challenges by adapting knowledge of state-of-the-art networks such as AlexNet, GoogleNet, OverFeat, etc. However, these networks have high number of parameters. This research proposes a five-layer architecture which has fewer parameters compared with above state-of-the-art networks, and can be also complementary to other convolutional neural network features. Extensive experiments on UC Merced and WHU-RS datasets prove that although our network decreases the number of parameters dramatically, it generates more accurate results than AlexNet, OverFeat, and its accuracy is comparable with other state-of-the-art methods.

Keywords: Satellite image classification, convolutional neural network, feature fusion.

1. Introduction

DUE to the invention of imaging devices like hyper spectral sensors, synthetic aperture radar, airborne visible/infrared Imaging spectrometer (AVIRIS) etc., more and more instruments are being developed to facilitate earth observation that allow us to examine the ground surface in greater detail. However, the very high resolution scenes images, inter-class similarity among scene categories or intra-class variability, significantly effects the classification performance. Although, scene categories are different from each other, the differences are almost undistinguishable due to identical thematic classes.

For example, images from forest and park, which belong to two different scene categories, may both consist of trees, mountains, and water at the same time but differ in the density and spatial distribution of these three thematic classes [1]. In this regard, existing approaches can be classified into three main components [2], namely: local visual methods, global visual methods and methods based on high-level vision information. Low-level features are handcrafted features, usually consist of color, shape or textual information [3], [4]. Although these features have been utilized effectively in different applications, an object or pixel-based information cannot fulfill the entire scene understanding due to its high-diversity and the various thematic classes. Mid-level features are potentially more distinctive than the traditional low-level local features and attempt to produce a global scene representation based on handcrafted features [5], [6]. One of the famous approach is the bag-of-words (BoW) model [7], but the challenge that comes along is the semantic gap between low-level visual features and high level semantics of images. BoW generally Neglects the context information between local patches. In this regard, the focus is shifting to high-level semantic features rather than depending on low-level or mid-level features.

A substantial progress has been made for acquiring high level semantic features due to the evolution of convolutional neural network (CNN) [8] because it encodes spectral and spatial information based on stack of convolutional filters. It has been utilized in many industrial applications and academic research in recent years as it continues to advance technologies in areas, like face retrieval [9], image segmentation [2], or even in the gaming world [10]. One approach is to construct a CNN from the start as a deep feature extractor. Another approach is to fine-tune the parameters by using a pre-trained CNN model that has been pre-trained on a subset of the ImageNet database, which is known as transfer learning.

Fig. 1. Example of photo gallery datasets.

Existing CNNs like AlexNet, GoogleNet, ResNet, and VGGNet have 57M, 6M, 24M and 138M parameters, respectively [11]. However, a high amount of memory is required to implement these pre-trained CNNs. Indeed, this issue is addressed by introducing a small parameters (204K) architecture for the food image recognition [11], [12]. However, there is no significant work that introduces small architecture for the remote-sensing scene classification. Therefore, we construct a five-layer CNN, which consists of 238K parameters, a bit higher than [11]. Example of proposed dataset is described in Fig. 1. We claim that proposed model can be also beneficial to other CNNs features. For this, a feature fusion strategy based on canonical correlation analysis (CCA) [13] is explored. The main contributions of this article are summarized as follows:

- A five-layer convolutional neural network is constructed for the remote-sensing scene image classification.
- The proposed CNN has less parameter compared with state-of-the-art networks such as AlexNet, OverFeat, etc.
- It can be complementary to other state-of-the-art CNNs features to improve the classification performance.

This paper is organized as follows. In Section II, the proposed architecture and a brief introduction of canonical correlation analysis is described. In Section III, a description of two datasets and experimental setup is presented. A comparative assessment of proposed CNN with baseline approaches is made in section IV. Finally, we draw conclusions for this paper in section V.

2. Proposed Architecture

The convolutional neural network used in this study is schematically illustrated in Fig. 3. It consists of four layers of hidden neurons that have learnable weights and biases (three convolutional-pooling and one fully-connected), apart from the last layer holds the output (the input is not regarded as a layer). The size of comprising RGB input is 32×32×3.

Each convolutional layer is the core building block and contains a set of learnable filters. The local receptive field of the first convolutional-pooling layer is 3 × 3 with a stride length of 1 pixel to extract 16 feature maps. A max pooling layer operation is performed for down-sampling in a 2 × 2 region. A channel-wise local response (cross channel) normalization layer is used for the first convolutional layer.

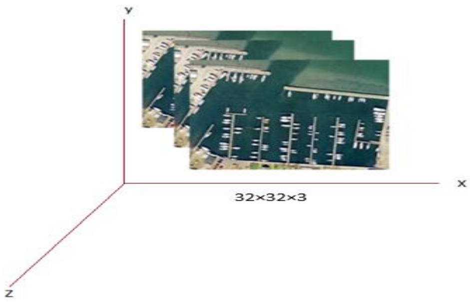

Fig. 2. Example of remote sensing 3D image.

For the second and third convolutional-pooling layer, we keep the same kernel size (3 × 3) to maintain the parameters value small, but the feature maps are 32 and 64, respectively. Instead of max pooling, an average pooling is performed for the third convolutional layer, and the other parameters remain unchanged. Batch normalization layers and the ReLU activation functions are used in each layers to speed up the training process. The fully-connected layer is the fourth layer which is implemented with weight learn factor (20) and bias learn factor (20).

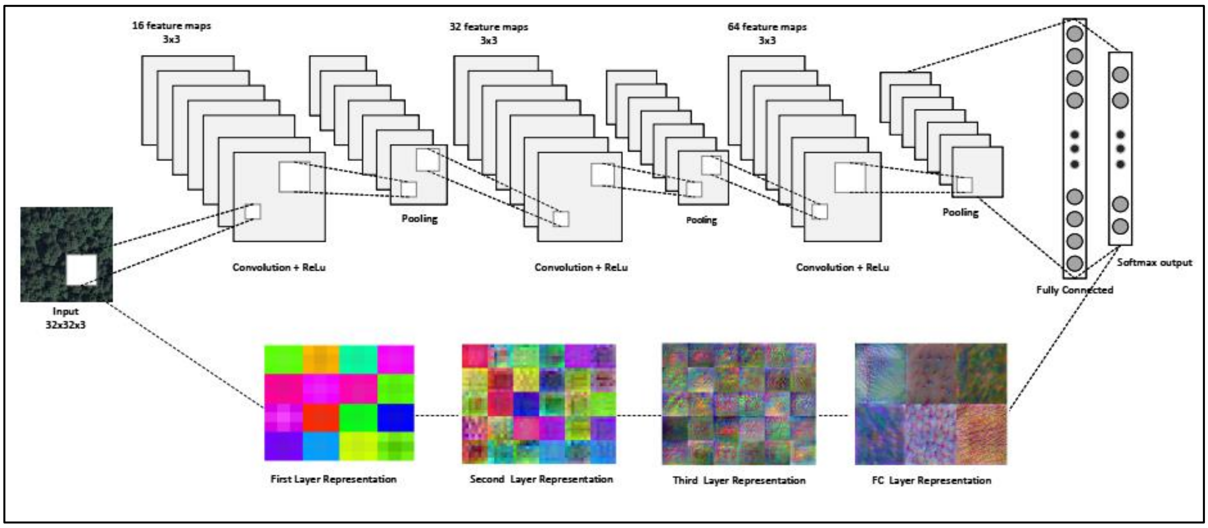

Fig. 3. The proposed architecture with feature visualization.

An increase in weight helped to increase the accuracy from 1% to 2%. The output layer has 19 softmax neurons that correspond to the nineteen categories of WHU-RS dataset. The stochastic gradient descent algorithm with a cross entropy cost function is used for training the proposed network. A mini-batch with 64 observations at each iteration is conducted and the learning rate by a factor of 0.2 every 8 epochs. The dropout layer is placed before the fully-connected layer to reduce over fitting. Training neural network requires learning the parameters and finding good hyper parameters, among which performing a gradient check is a simple gradient. To take into consideration, a relative error is set to 1e-4 through trials and errors.

2.1. Canonical Correlation Analysis

The canonical correlation analysis (CCA) is one of the statistical method that establish the correlation criterion function between two groups of feature vectors, to extract their canonical correlation features. In our work, it performed to fuse the fully-connected layer features of proposed CNN with AlexNet and VGGNet-16. Here $X \in R^{p\times n}$ and $Y \in R^{q\times n}$ represent two matrices, each consists of n training feature vectors from two different sets, p and q are the dimensions of each vector.

Assume that $S_{xx} \in R^{p\times p}$ and $S_{yy} \in R^{q\times q}$ contain within-sets covariance matrices of X and Y, $S_{xy} \in R_{p\times q}$ contains the between-set covariance matrix (consider as $S_{yx} = S^T_{xy}$. The overall covariance matrix $(p + q) \times (p + q)$ is then computed as

$$S = \begin{pmatrix} cov(x) & cov(x,y) \\ cov(y,x) & cov(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{1}$$

It is complicated as stated in [32] to follow the relationship between these two sets of vector from matrix S because these feature vectors may not follow a consistent pattern. The objective of CCA is to find the linear combinations, $\overset{*}{X} = \mathrm{W}_x^T\mathrm{X}$ and $\overset{*}{Y} = \mathrm{W}_y^T\mathrm{X}$ which maximizes the pair-wise correlations across the two feature sets:

$$\mathrm{corr}\begin{pmatrix} \overset{*}{X} & \overset{*}{Y} \end{pmatrix} = \frac{cov\begin{pmatrix} \overset{*}{X} & \overset{*}{Y} \end{pmatrix}}{Var(\overset{*}{X}).var(\overset{*}{Y})} \tag{2}$$

where $\mathrm{cov}\begin{pmatrix} \overset{*}{X} & \overset{*}{Y} \end{pmatrix} = \mathrm{W}_x^T \mathrm{SxyWy}$,

$$\mathrm{var}(\overset{*}{X}) = W_x^T \mathrm{SxxWx},\ \mathrm{var}(\overset{*}{Y}) = \mathrm{W}_Y^T \mathrm{SyyWy}$$

Maximization is conducted by maximizing the covariance between $\overset{*}{X}$ and $\overset{*}{Y}$ using Lagrange multipliers subject to satisfy the following constraintsvar$(\overset{*}{X}) = \mathrm{var}(\overset{*}{Y}) = 1$. Both transformation matrices, W_x and W_y, are then computed by using the eigenvalue equations:

$$\begin{cases} S_{xx}^{-1}\, Sxy\, S_{yy}^{-1}\, Syx\widehat{W} = R^2\widehat{W_x} \\ S_{yy}^{-1}\, Syx\, S_{xx}^{-1}\, Sxy\widehat{W} = R^2\widehat{W_y} \end{cases} \tag{3}$$

where $\widehat{W_x}$ and $\widehat{W_y}$ are the eigenvectors and R^2 is the diagonal matrix of eigenvalues or it could be defined as squares of the canonical correlations. The number of non-zero eigenvalues can be find in each equation, that is $d = \mathrm{rank}\,(\mathrm{Sxy} \le \min(\mathrm{n,p,q})$, which will be fixed in descending order, $r1 \ge r1 \ge \cdots \ge rd$. As mentioned earlier, both the transformation matrices, W_x and W_y, Composed of the sorted eigenvectors corresponding to the non-zero eigenvalues. $\overset{*}{X}, \overset{*}{y} \ \epsilon\ R^{d\times n}$ are consider as canonical variates. It could be observed that the sample covariance matrix denoted in Eq. (1) will be of the form:

$$\overset{*}{S} = \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & r_1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & r_4 \\ \hline r_1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_4 & 0 & 0 & \cdots & 1 \end{array}\right) \tag{4}$$

The matrix explains that the canonical variates have nonzero correlation only on their corresponding indices. It also express that the canonical variates are uncorrelated within each other because of identity matrices in the upper left and lower right corners. Hence, it is possible to perform feature-level fusion either by concatenation or summation of the transformed feature vectors:

$$Z_1 = \begin{pmatrix} \overset{*}{x} \\ \overset{*}{x} \end{pmatrix} = \begin{pmatrix} W_x^T\ X \\ W_y^T\ X \end{pmatrix} = \begin{pmatrix} W_x & W_y \\ 0 & 0 \end{pmatrix}^T = \begin{pmatrix} x \\ y \end{pmatrix} \tag{5}$$

Or

$$Z_2 = \overset{*}{x} + \overset{*}{y} = W_x^T\ X + W_y^T\ Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{6}$$

Here Z_1 and Z_2 are Canonical Correlation Discriminant Features (CCDFs).

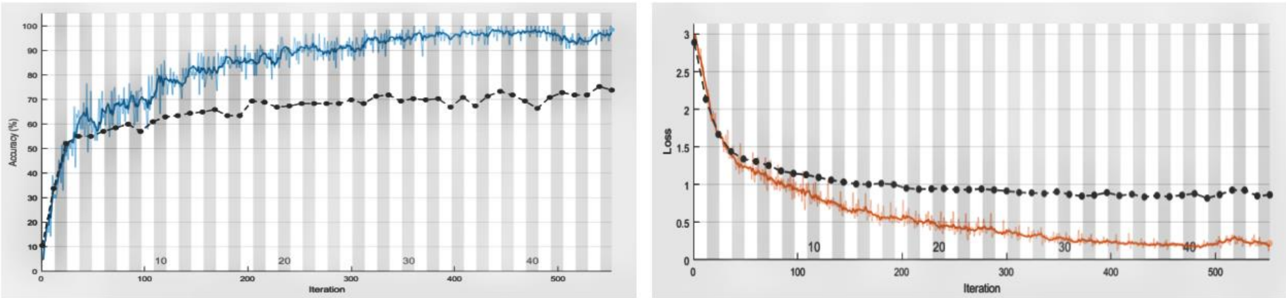

Fig. 4(A). Accuracy (left) and loss (right) curves with training process is performed without data augmentation.

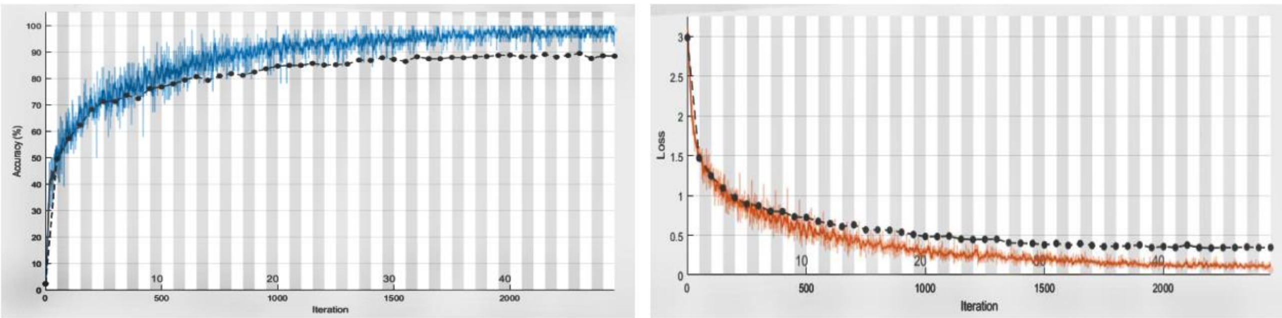

Fig. 4(B). Accuracy (left) and loss (right) curves with training process is performed with data augmentation.

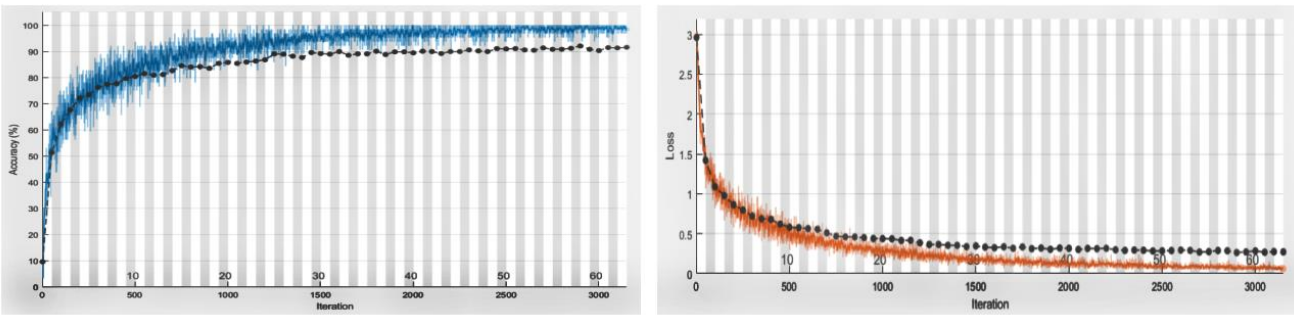

Fig. 5(A). Accuracy (left) and loss (right) curves for whu-rs dataset. The training process was stop since no increase in accuracy after 65 epochs.

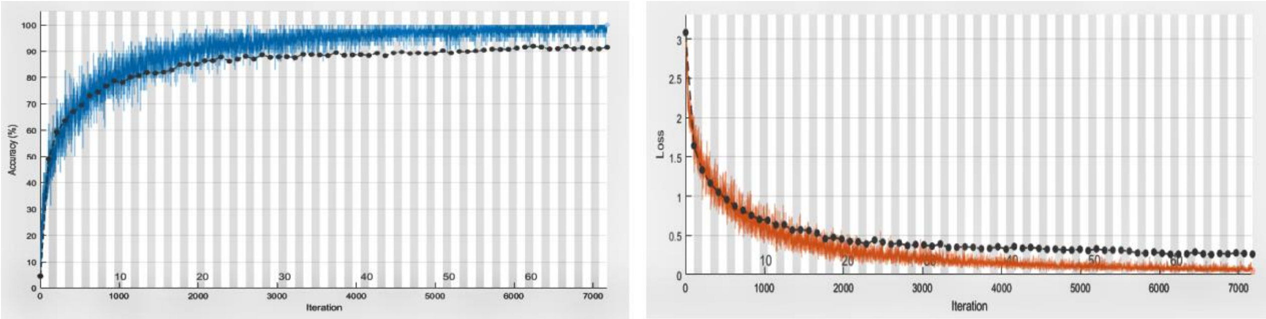

Fig. 5(B). Accuracy (left) and loss (right) curves for uc-merced datasets. the training process was stop since no increase in accuracy after 70 epochs.

3. Results and Discussion

The first dataset for evaluating the performance of the proposed CNN is the widely used UC Merced, which was acquired from the USGS National Map Urban Area Imagery collection, contains 21 distinctive scene categories (100 samples each). The image size is 256 × 256 pixels with a pixel resolution of one foot. The second data set, WHU-RS [14], consists of 950 images (19 scene categories) with a size of 600 × 600 pixels.

The CNN was first trained with the maximum training epoch set to 50. (Fig. 4(a)) shows the accuracy and loss curves of WHU-RS dataset. There is a large gap between training and test appeared after 8 epochs, showing the presence of over fitting. In order to deal with this issue, a data augmentation strategy is proposed. The authors in [15], introduce spatial pyramid pooling that allows CNNs to use multi-scale images. Inspired by their work, we also expand images by using affine transformation. This simple data augmentation is adopted by incorporating three types of transformations including, rotation, translation, and scaling to all the images. Rotation is done by rotating the images 40 degree clockwise using bilinear interpolation. Translation is performed by using the translation vector [80 80]. Images were scaled using factor 5 by bicubic interpolation.

Then, the CNN was trained with the data augmentation strategy. As illustrated in (Fig. 4(b)), two issues were solved through almost trivial data transformations. First, the classification performance is greatly lifted to a level up to 90% within 50 epochs, and second, the issue of over fitting is well addressed, and eliminated. From the Fig. 2(b), it seemed that further improvement is possible by using more training epochs.

So, the CNN is trained with another training scheme (80 epochs). (Fig. 5(a)) and (Fig. 5(b)) show accuracy and loss curves of WHU-RS dataset and UC Merced dataset, respectively. The accuracy did increase further, but stopped after 70 epochs. Hence, training the CNN with 70 epochs resulted in the highest accuracy of more than 93% for WHU-RS dataset and 92% for UC Merced dataset. The problem of over fitting could be occurred if we keep increasing training epochs without further improvement in accuracy.

To claim that proposed CNN can be complementary to other networks, the pre-trained AlexNet [34] and VGGNet-16[16] are employed to extract deep features and their fully connected (FC) layer features were regarded as separated feature descriptors. Then, the FC features of proposed CNN were fused with second fully-connected layer features of AlexNet and VGGNet by concatenation. Table 2 illustrates the performance after fusion. As stated in[17], a simple addition of the features can increase accuracy from 1% to 2%. The proposed fusion strategy increases the classification performance 4% to 5% compared to their pure features. It bears some resemblance to PCA, but PCA searches for patterns within a single multivariate dataset while CCA identifies new variables that maximize the inter-relationships between two datasets.

4. Comparison with The State-of-the-Art Methods

To make a fair comparison with other methods, first, we compare the performance of proposed CNN without using feature fusion strategy. As shown in Fig. 4(b)and illustrated in Table 1, the highest classification rate for the UC Merced dataset is 92.80%, which is higher than baseline methods such as AlexNet [18], OverFeat [19], and comparable to GoogLeNet [20].

Table 1. Overall Classification Accuracy (%) of Reference and Proposed Methods on the UC-Merced Dataset and WHU-RS Dataset

Method	UC Merced
LPCNN[21]	89.90%
CCNN[22]	91.56%
S-UFL[23]	82.72±1.18%
SRSCNN-NV[22]	92.58%
GoogLeNet[20]	92.80±0.61%
AlexNet[18]	90.21±1.17%
SPP-net+SV[22]	91.38±0.46%
OverFeat[19]	90.91±1.19%
Proposed CNN	92.80±0.85%

Other state-of-the-art methods including, a large patch convolutional neural network (LPCNN), where authors replace the fully-connected layer with global average pooling layer to decrease the total parameters [21], an unsupervised feature learning approach to extract patches based on saliency detection algorithm [28], a new convolutional neural network for dealing the scale variation of the objects in the scenes [22], and a deep CNN with spatial pyramid pooling (SPP-net) to extract multi-scale deep features. The proposed CNN architecture achieves very competitive accuracy in the literature of scene classification when compared with these state-of-the art approaches. For further analysis, a confusion matrix of UC Merced dataset and WHU-RS dataset is shown in Fig. 4, and Fig. 5, respectively. The tennis court category in UC Merced dataset, which is hard to be classified because of inter-class similarity with golf course, achieving lower accuracy. From the confusion matrix of WHU-RS dataset, it seems that other categories such as railway station, pond, park, forest, commercial, and airport are easily confused due to similar structures and background color. In summary, these datasets are challenging, even though we have achieved a comparable performance.

Table 2. Overall Classification Accuracy (%) of Reference and Proposed Methods on the UC-Merced Dataset and WHU-RS Dataset

Methods	UC-Merced	WHU-RS
GoogLeNet+ fine-tuning [24]	97.10%	96.14%
GBRCN [7]	94.53%	-
UFL [12]	95.71%	-
MARTA GANs [25]	94.86±0.80%	-
D-CNN with AlexNet [2]	96.67±0.10%	-
D-CNN with GoogLeNet[2]	97.07±0.12%	-
D-DSML-CaffeNet[26]	96.76±0.36%	96.64±0.68%
Fusion by addition[27]	97.42±1.79%	98.70±0.22%
CaffeNet[14]	95.02±0.81%	96.24±0.56%
VGG-VD-16 [14]	95.21±1.20%	96.05±0.91%
AlexNet-SPP-SS[18]	96.67±0.94%	95.00±1.12%

SPP-net+MKL [22]	96.38±0.92%	95.07±0.79%
Fusion with AlexNet	97.73±0.80%	98.47±0.40%
Fusion with VGGNet-16	98.40±0.30%	99.23±0.50%

To demonstrate the performance of proposed fusion strategy (CCA), we compare the results with other state-of-the art methods as illustrated in Table 2. In [27], discriminant correlation analysis (DCA) is proposed to fuse the two fully connected layer features of VGG-Net architecture. The work reported in[14], attempts to tune the weights of CaffeNet using fine-tuning approach based on VGG-VD-16 architecture. A pre-trained Alex-Net is used in [13], with spatial pyramid pooling (SPP-net), and to fuse the multi-layer features, the multi-kernel learning is proposed.

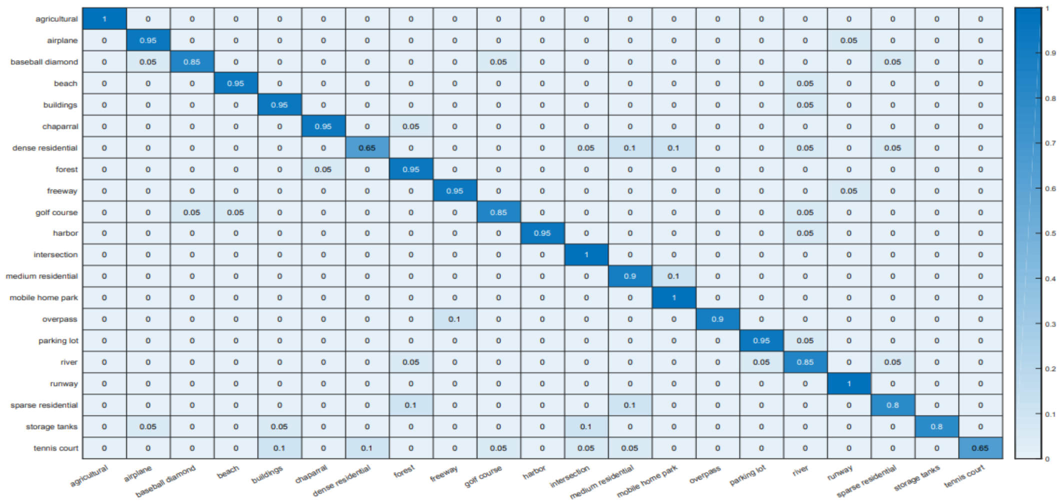

Fig. 6. Confusion matrix for the UC-merced dataset.

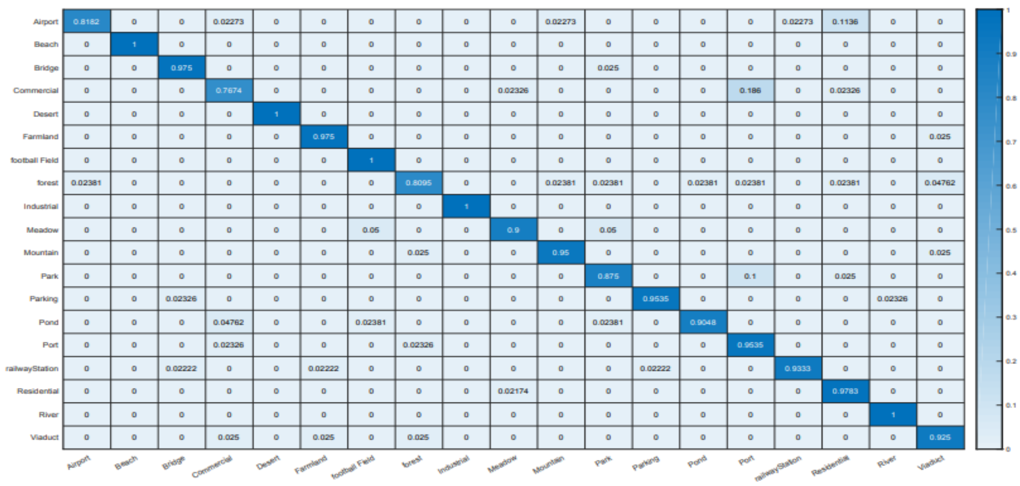

Fig. 7. Confusion matrix for the WHU-RS dataset.

Metric learning (ML) [2], [26], has been utilized frequently to further increase the discrimination of deep representations. To increase the depth of convolutional layers, the side supervision strategy (SS) is proposed for AlexNet model [18]. Other approaches including, an unsupervised representation for deconvolution networks [12], [25], fine-tuned GoogleNet [24], fine-tuned CaffeNet with VGG-VD-16 [14], and a gradient boosting random CNN [23]. Compared to these methods, proposed fusion strategy with VGGNet-16 achieves the best accuracy (99.23%) for WHURS dataset using 80% samples as training data and obtained an impressive accuracy (98.40%) for UC Merced dataset.

5. Conclusion

Existing works pay little attention in constructing a small network. High amount of memory is required if we implement the models such as AlexNet, GoogleNet, Inception etc. Therefore, we attempt to fill this gap and propose a five-layer CNN model to achieve a competitive accuracy for the remote sensing scene image classification. During the training, over fitting issue was also well-addressed and eliminated through affine transformations. The incorporation of a fusion strategy leads to an encouraging results. However, it does not take account the real-time requirements, but well suited for off-line classification, where the classification accuracy is the prime goal. The feature fusion also reduce the diverseness of feature representation. In the future work, we would like to focus on these challenges.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Muhammad Ashad Baloch write code, implement ideas and methodology. Sajid Ali gave me the idea about this topic and support me for acquiring this task. Mubashir H.Malik contribute in preprocessing Step writing. Amir Hussain provided the facility of lab for conducting the experimental work. He also supported me in my experimental work. Abdul Mustaan Madni helped to draw all diagrams and to write introduction section.

Acknowledgement

The Authors heartily acknowledge the Dr. Sajid Ali, Faculty, Department of Computer science Education University Lahore for his cooperation and supervision for acquiring this task and also providing the lab facilities for conducting the experimental work.

References

[1] Hu, F., *et al.* (2016). Fast binary coding for the scene classification of high-resolution remote sensing imagery*. Remote Sensing, 8(7),* 555.

[2] Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE, 105(10),* 1865-1883.

[3] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60(2)*, 91-110.

[4] Ahonen, T., *et al.* (2004). Face recognition based on the appearance of local regions. *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004.*

[5] Qi, K., *et al*. (2016). Multi-task joint sparse and low-rank representation for the scene classification of high-resolution remote sensing image. *Remote Sensing, 9(1)*, 10.

[6] Zhao, B., *et al.* (2016). The fisher kernel coding framework for high spatial resolution scene classification. *Remote Sensing, 8(2),* 157.

[7] Jiang, Y. G., Ngo, C. W., & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

[9] Dong, Z., *et al.* (2018). Deep CNN based binary hash video representations for face retrieval. *Pattern Recognition, 81,* 357-369.

[10] Andersen, P. A., Goodwin, M., & Granmo, O. C. (2018). *FlashRL: A Reinforcement Learning Platform for Flash Games.*

[11] Heravi, E. J., Aghdam, H. H., & Puig, D. (2018). An optimized convolutional neural network with bottleneck and spatial pyramid pooling layers for classification of foods. *Pattern Recognition Letters, 105,* 50-58.

[12] Lu, Y. (2016). *Food Image Recognition by Using Convolutional Neural Networks (CNNs).*

[13] Sun, Q. S., *et al.* (2005). A new method of feature fusion and its application in image recognition. *Pattern Recognition, 38(12),* 2437-2448.

[14] Xia, G. S., *et al.* (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 55(7),* 3965-3981.

[15] He, K., *et al.* (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *Proceedings of European Conference on Computer Vision*. Springer.

[16] Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition. 1556.*

[17] Li, E., *et al.* (2017). Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 55(10),* 5653-5665.

[18] Han, X., *et al.* (2017). Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing, 9(8),* 848.

[19] Penatti, O. A., Nogueira, K., & dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[20] Nogueira, K., Penatti, O. A., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition, 61,* 539-556.

[21] Zhong, Y., Fei, F., & Zhang, L. (2016). Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing, 10(2),* 025006.

[22] Liu, Q., *et al.* (2016). *Learning multi-scale deep features for high-resolution satellite image classification.*

[23] Zhang, F., Du, B., & Zhang, L. (2015). Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 53(4),* 2175-2184.

[24] Castelluccio, M., *et al.* (2017). Training convolutional neural networks for semantic classification of remote sensing imagery. *Urban Remote Sensing Event (JURSE),* IEEE.

[25] Lin, D., *et al.* (2017). MARTA GANs: Unsupervised representation learning for remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters, 14(11),* 2092-2096.

[26] Gong, Z., *et al.* (2018). Diversity-promoting deep structural metric learning for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 56(1),* 371-390.

[27] Chaib, S., *et al.* (2017). Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens, 55(8),* 4775-4784.

[28] Zhang, F., Du, B., & Zhang, L. (2016). Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing, 54(3),* 1793-1802.

Muhammad Ashad Baloch was born in Multan, Pakistan in 1994. He did M.Phil. computer science degree from National College of Business Administration & Economics Lahore (Multan Campus), Pakistan in 2019. Now he is working as a lecturer in computer science at University of Education Lahore, Pakistan. This research is a part of M.Phil. thesis

"An Efficient Convolutional Neural Network for Remote-Sensing Scene Image Classification".

Sajid Ali received the MSc (CS) degree in 2003 and the MS (CS) in 2005 in computer science from Department of Computer Science, the Agriculture University, Faisalabad, Pakistan, respectively. He received PhD and postdoctoral (CS**)** from Beijing Normal University, Beijing, China in 2013 and 2015, respectively. Currently, he is a faculty member at the Department of Information Sciences, University of Education, Lahore, Pakistan. His current research interests include motion sensor, 3D-human motion, biometrics technology, animation, digital image processing and information system, computer network.

Mubasher H. Malik is working as an assistant professor in the Department of Computer Science. His area of interest for research is computer vision, image processing, machine learning. From a decade, he is accelerating his research journey in the Area of artificial intelligence.

Aamir Hussain received the Ph.D degree in computer science and technology from the School of Computer Science and Technology, Wuhan University of Technology, China, in 2016. He is currently an assistant professor with the Department of Computer Science, Muhammad Nawaz Shareef University of Agriculture Multan, Pakistan. His research interests include wireless sensor networks, internet of things, and Software-Defined Networks (SDN).

Abdul Mustaan Madni was born in Dera Ghazi khan, Pakistan in 1995. He is doing M.Phil. computer science degree from National College of Business Administration & Economics Lahore (Multan Campus), Pakistan. Now he is working as a developer in PROFEXEO software house. His research interests are in data science.