Fusion of Deep Learning Models for Improving Classification Accuracy of Remote Sensing Images

P. Deepan 1, L.R. Sudha 2

1 Research Scholar, Department of Computer & Science and Engineering, Annamalai University, Chidambaran, Tamil Nadu, India
2 Associate Professor, Department of Computer & Science and Engineering, Annamalai University, Chidambaran, Tamil Nadu, India

1 deepanp87@gmail.com, 2 sudhaselvin@ymail.com

Corresponding Author: P. Deepan

Corresponding Author: P. Deepan

Abstract

Over the recent years we have witnessed an increasing number of applications using deep learning techniques such as Convolutional Neural networks (CNNs), Recurrent Neural Networks (RNN) and Deep Neural Networks (DNN) for remote sensing image classification. But, we found that these models suffer for characterizing complex patterns in remote sensing imagery because of small inter class variations and large intra class variations. The intent of this paper is to study the effect of ensemble classifier constructed by combining three Deep Convolutional Neural Networks (DCNN) namely; CNN, VGG-16 and Res Inception models by using average feature fusion techniques. The proposed approach is validated with 7,000 remote sensing images from Northern Western Polytechnical University – Remote Sensing Image Scene Classification (NWPU- RESISC) 45 class dataset and confirmed as an effective technique to improve the robustness over a single deep learning model.

Keywords: Image classification, Remote sensing, Feature fusion, Convolutional neural network, Deep CNN and Ensemble classifier.

I. Introduction

Remote sensing images play an important role in a wide range of applications such as image classification, object detection, land use/land cover analysis, change detection and found to be useful in various applications such as geographic information system (GIS) update, environmental monitoring, geological hazard detection, Land Use/Land Cover (LULC) mapping, precision agriculture and urban planning [III]. Remarkable efforts have been made to develop various methods for...
Object classification methods are divided into three categories based on the features they use, namely handcraft feature learning method, unsupervised feature learning method and deep feature learning based method [XXIV]. Earlier, scene classification was based on the handcraft feature learning based method. This method [XV] was mainly used for designing the engineering features, such as colour, shape, texture, spatial and spectral information. The unsupervised feature learning method [XXIII] is an alternate for the handcrafted feature method and learning the unlabeled data for remote sensing image classification. The aim of unsupervised feature learning method [II] is used to identify the low-dimensional features that capture some underlying high-dimensional input data. When the feature learning is performed in unsupervised way, it enables a form of semi-supervised learning where features learned from an unlabeled dataset are then employed to improve performance in a supervised setting with labelled data. There are several unsupervised feature learning methods available such as k-means clustering, principal component analysis (PCA), sparse coding and auto encoder. In real time applications, the unsupervised feature learning methods have achieved high performance for classification compared to handcrafted-feature learning methods. However, the lack of semantic information provided by the category label cannot promise the best discrimination between the classes. So we need to improve the classification performance and extract powerful discriminant features for improving classification performance.

The deep learning model is composed of multiple processing layers that can learn more powerful feature representations of data with multiple levels of abstraction. When compared with handcrafted feature and unsupervised feature, the deep feature learning method has been found to automatically learn from data using deep neural networks. In deep feature learning method there are several number of learning models such as recurrent neural network (RNN), convolutional neural network (CNN), deep neural network (DNN), stacked auto encoder (SAE), and so on.

This paper has been organised as follows: Section 2 covers literature survey of the proposed approach. Section 3 includes the architecture of proposed work. Section 4 provides description about the benchmark dataset. Section 5 discusses about the experimental result and analysis. Finally conclusions are given in Section 6.
II. Related Works

Earlier, image classification methods were based on traditional machine learning classification such as Support Vector Machine (SVM), k-Nearest-Neighbour (kNN), Regression, Decision Tree and Random Forest. Mountrakis et al. [XIII] discussed about the applications of SVMs in remote sensing. In many cases, the SVM classifiers have better accuracy, stability and robustness when compared with other classifiers such as neural network and k-nearest neighbour. Thanh Noi.P et al. [XIX] presented an approach that compared land cover classification of remote sensing images with non-parametric classifier such as SVM, k-NN and random forest algorithms. In their work, they have chosen fourteen classes of data compared with the above mentioned three classifiers. Ayhan et al. [I] analyzed various image classification methods for remote sensing images. In this study, researchers have compared artificial neural networks, standard maximum likelihood classifier and fuzzy logic method. Based on the comparison, ANN classification is more robust than other two classifiers. Comparison between the k-NN and random forest were presented in a study by McInerney et al. [XII] where, remote sensing classifications along with performance metrics are carried out. As, the current scale of images accessible on the internet is massive, it is difficult to handle large image dataset with traditional classification methods. So in order to handle the large image set deep learning based object classification methods are introduced. Recently, many deep learning based object classification techniques have been proposed to solve the classification problem. One of them is Convolutional Neural Networks (CNNs).

Deep learning is a powerful technique that can be widely used in various visual processing tasks including remote sensing images. In particular, Convolutional neural networks (CNN) and Deep Convolutional Neural Networks (DCNN) obtain the modern results for automated object detection and classification in remote sensing imagery. The CNN model can be used to classify the remote sensing imagery, such as satellite image, high resolution imagery, multispectral and hyperspectral imagery. Maher Ibrahim Sameen et al. [X] proposed classification of Very High Resolution (VHR) images using spectral-spatial CNN. Maarten C. Kruithof et al. [IX] proposed object recognition based DCNN model with complete transfer and partial frozen layers. Yang Long et al. [XXII] developed a new object based localization framework that can be classified the data into three categories namely; region proposal, classification and process of object localization. Shuying Liu et al. [XVI] introduced DCNN model for image classification for small training data. Krizhevsky et al. [VII] trained and classified the Image Net dataset by using Deep Convolutional Neural Networks.

However, CNN model requires huge amount of training data, which rise major challenge in the remote sensing field. The researchers have to solve the training data
lacking problems by introducing data augmentation techniques for training data. Xingrui Yu et al. [XXI] proposed a method to enhance the volume of training data in the remote sensing imagery. The enhancement is based on applying flip, translation and rotation operation for the original dataset. Luke Taylor et al. [XVIII] introduced a data augmentation technique, which provides artificial transformation to the original training data.

Now a days, ensemble classification or feature fusion model is broadly used for image classification from remote sensing imagery. The aim is to combine the results of two or more individual models. Ensemble classification is also called as committee based learning. It’s an effective method to develop accurate classification system. The ensemble model is able to boost weak learners that are slightly better than the random guess to strong aggregated learners which can make good accurate predictions. Moreover, the generalization ability of an ensemble classifier is much stronger than base learners. Ensemble is consisting of two steps namely, generating the base learning and then combining the base learners. Travis Williams et al. [XX] introduced ensemble based image classification by using wavelet transform. This model convert the data into wavelet domain to achieve better accuracy and efficiency for image classification. Grant J. Scott et al. [VII] proposed a new ensemble method for image classification based on pre-trained Deep Convolutional Neural Network (DCNN) models. Souleyman Chaib et al. [XVII] proposed a deep feature fusion for very high resolution remote sensing image classification by fusion of VGG-16 model and Alex Net model. This fusion model is experimented with three benchmark datasets namely, UC MERCEd data set, WHU-RS data set and Aerial Image Dataset (AID). Dimitrios Marmanis et al. [XXIII] introduced a novel feature fusion algorithm, which provides better accuracy, efficient result and also tackled huge amount of training data. Muhammet Ali Dede et al. [XI] proposed a fusion feature based image classification by using the Dense Net and Inception networks model. Zhiling Guo et al. [XXV] proposed a tool for identifying village buildings based on the Ensemble based CNN model. Cheng Ju et al. [V] investigated and improved the efficiency of DCNN by using various ensemble methods like voting rule. The result of feature fusion model is more accurate rather than individual DCNN model.

III. Proposed Work

In this paper, we focus on the ensemble model for improving the performance and increasing the predication rate of remote sensing image classification. We have introduced three standard Convolutional Neural Networks such as CNN, VGG-16 and Res Inception as our base models. Each base model is trained independently with 10 class dataset after finetuning the model parameters such as batch size, activation function and Dropout Probability. Then we have calculated the average of all the features after fully connected layers and applied softmax function to classify the data.
The architecture diagram of the proposed method is shown in Fig. 1 and the details of the components in the architecture diagram are discussed in the following sections.

![Architecture of average feature fusion model](image1)

**Convolutional Neural Network**

The CNN model consists of four hidden layers (three convolutional-sub sampling layer pair and one fully connected layer), one input layer and one output layer. The input layer contains 224x224x3 neurons, indicating the RGB values for a 224x224x3 image. The convolutional-sub sampling layer use the size 3x3 stride lengths followed by 2x2 regions. The number of filter is gradually increased from size of 64, 128 and 256. The fully connected layer contains 1024 neurons with Rectified Linear Unit (ReLU). Finally, the output layers consist of 10 class object classification as output by using softmax function. The Fig. 2 represents the working flow of convolutional neural networks.

![Architecture of Convolutional Neural Network Model](image2)

**VGG-16 network**

VGG Net is abbreviation for Visual Geometry Group Networks. The model is one of the most powerful deep convolutional neural networks which was proposed by Simonyan et al. It consists of 13 convolutional layers with 3x3 filter size, 5 sub...
sampling/ max pooling layer with size of 2x2, two fully connected layers with activation function and softmax function. Each block is made by consecutive 3 × 3 convolutions and followed by a max pooling layer. The number of filter is gradually increased from size of 64, 128, 256, 512 and 512. To avoid the problem of over-fitting, we need to eliminate the redundant features by adding the Dropout. We propose the VGG-16 Net, which is efficient model for classifying the remote sensing images. The depth of this architecture makes it suitable for extracting the deep features from the data. Fig. 3 shows the architecture of VGG-16 model.

![Fig. 3 Architecture of VGG-16 Model](image)

### Inception Residual Network

The model is hybrid version of Inception and Res Network. For the residual versions of the Inception networks, we use cheaper Inception blocks than the original Inception. Each Inception block is followed by filter-expansion layer (1 × 1 convolution without activation) which is used for scaling up the dimensionality of the filter bank before the residual addition to match the depth of the input. This is needed to compensate for the dimensionality reduction induced by the Inception block. Another small technical difference between our residual and non-residual Inception variants is that in our Res Inception model experiments, we have used batch-
normalization only on top of the traditional layers, but not on top of the residual summations. It is reasonable to expect that a thorough use of batch-normalization should be advantageous. But, the implementation of batch-normalization in TensorFlow was consuming a lot of memory and we would have needed to decrease the overall number of layers, if batch-normalization had been used everywhere.

**Average Feature Fusion Classifiers**

Averaging feature fusion technique is one of the familiar ensemble approaches for classifying remote sensing images. It takes average of output score divided by probability of all base CNN model and report it as predicted score divided by probability. Due to high capacity of Deep CNN, the average feature fusion model improves the performance substantively. Taking average of multiple models will reduce the variance.

**IV. Dataset Description**

Readily available benchmark dataset NWPU-RESISC 45 class has been used for Remote Sensing Image object classification [III]. This dataset is created by the North Western Polytechnical University (NWPU) from Google Earth. The dataset contains 31,500 satellite images and each class consists of 700 images. The resolution of each image has a size of 256x256 pixels with RGB colour space. In our proposed method, we have randomly selected ten classes such as airplane, beach, bridge, desert, forest, lake, river, storage tank, tennis court and wetland. The Fig. 4 shows some example images from NWPU-RESISC 45 class dataset for image classification.

![Fig. 4 Dataset description of NWPU-RESISC 45 class dataset](image)

In Deep CNN model, the dataset has been split into training, validation and testing datasets separately. The validation images have been randomly selected from training
sample based on the size of validation. The testing images are used to predict the result based on training datasets.

Performance Evaluation Metrics

We have to evaluate the performance of a proposed model by using various performance metrics such as Precision, Recall, Accuracy, F1-measure and Mean Square Error (MSE). The confusion matrix is a two dimensional table, which is used to calculate the above mentioned metrics. In this matrix actual classifications are in column side and predicted values are in row side. The Fig. 5 shows the confusion matrix table. Let TP, TN, FP and FN denote the number of true positive, number of true negative, number of false positive and false negative respectively. The true positive is an outcome, where the models correctly predict the positive class. The true negative is an outcome, where the models correctly predict the negative class. The false positive is an outcome, where the models incorrectly predict the positive class. The false negative is an outcome, where the models incorrectly predict the negative class.

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>N</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Fig. 5 Confusion matrix

Precision

The precision measure can be calculated by number of true positive results divided by the number of positive results predicted by the classifier.

\[
P = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]  

Recall

The recall measure can be calculating the number of correct positive results divided by the number of all relevant samples.

\[
R = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]  

Accuracy
The accuracy measure can be calculated by dividing the number of correct predictions by the total number of input samples.

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} 
\]

(3)

**F1 -measure**

The F1-measure (harmonic mean) is used to show the balance between the precision and recall measures. The F1-score measure can be calculated as follow:

\[
F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} 
\]

(4)

**MSE**

The MSE value returns prediction error rate between the original and fused image. It is calculated by the equation below:

\[
\text{MSE} = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (O(i,j) - \hat{O}(i,j))^2}{H \times W} 
\]

(5)

Where, O and \(\hat{O}\) are represented as observed original image and feature fused image respectively and value of H and W represented as Height and Width of the images.

**V. Experimental Results and Analysis**

We have evaluated deep feature fusion CNN Model for classification of remote sensing image based on three CNN architecture pre-trained by Image Net. The input images are reshaped into 224x224 for pre-trained Deep CNN model. We found that Res Inception Model achieved 98% of accuracy and turned out to be the best result and VGG-16 model also produce nearly equal accuracy to the Res Inception model (94% of Accuracy). But, the CNN model achieved 65% of accuracy and turned out to be lowest classification result. The confusion matrix of three standard deep CNN models is shown in Fig. 6,7and 8. We have compared our proposed Average Feature Fusion (AFF) method with the above three well-known standard deep CNN model CNN, VGG-16 and Res Inception. The confusion matrix of proposed feature fusion model is shown in Fig. 9.
The proposed feature fusion model accuracy is nearly equal to Res Inception model. But, the Mean Square error rate is reduced from the Res Inception model. The Res Inception and feature fusion model MSE values are 0.26 and 0.05 respectively. The performance of remote sensing image classification for various pre-trained CNN models are shown in Table 1. Based on the result, AFF method outperformed than standard deep CNN models. Under the framework of tensorflow backend, the entire process was trained and run in Core i7 3.44GHz, 3TB of Hard disk and 16 GB of RAM. The comparative analysis of proposed model with standard DCNN architecture is shown in Fig. 10.
Table 1. Comparative analysis of proposed model performance

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Deep Learning Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CNN</td>
<td>0.65</td>
<td>0.71</td>
<td>0.65</td>
<td>0.65</td>
<td>5.44</td>
</tr>
<tr>
<td>2.</td>
<td>VGG-16</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.86</td>
</tr>
<tr>
<td>3.</td>
<td>Res Inception</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.26</td>
</tr>
<tr>
<td>4.</td>
<td>CNN, VGG-16</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>1.06</td>
</tr>
<tr>
<td>5.</td>
<td>CNN, Res Inception</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.24</td>
</tr>
<tr>
<td>6.</td>
<td>VGG-16, Res Inception</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.25</td>
</tr>
<tr>
<td>7.</td>
<td>Average Feature Fusion Method</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Fig. 10 Performance analysis comparison for various models

VI. Conclusion

In this paper, we have proposed an average feature fusion based Deep Convolutional Neural Networks model to improve the performance and prediction accuracy for object classification of remote sensing images. In order to demonstrate the efficiency of proposed approach experiments are conducted on a 10 class dataset selected from Northern Western Polytechnical University – Remote Sensing Image Scene Classification (NWPU- RESISC) using three distinct CNN models (CNN, VGG-16 and Res Inception) individually and found the accuracy as 0.65, 0.94 and 0.98 respectively. Then average feature fusion model is applied and achieved 99% of accuracy and also reduced the MSE rate up to 0.05. Implementing the proposed work
in GPU configuration will reduce the computational time of Remote Sensing Image Classification.

References


