Neural Networks 95 (2017) 19–28

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A patch-based convolutional neural network for remote sensing image classification

Atharva Sharma [a,*], Xiuwen Liu [a], Xiaojun Yang [b], Di Shi [c]

[a] Department of Computer Science, Florida State University, Tallahassee, FL 32306-4530, United States
[b] Department of Geography, Florida State University, Tallahassee, FL 32306-2190, United States
[c] Department of Geography and Atmospheric Science, University of Kansas, Lawrence, KS 66045-7316, United States

ARTICLE INFO

Article history:
Received 16 March 2017
Received in revised form 16 June 2017
Accepted 28 July 2017
Available online 8 August 2017

Keywords:
CNN
Deep learning
Remote sensing imagery
Medium-resolution
Spatial context
Patch-based

ABSTRACT

Availability of accurate land cover information over large areas is essential to the global environment sustainability; digital classification using medium-resolution remote sensing data would provide an effective method to generate the required land cover information. However, low accuracy of existing per-pixel based classification methods for medium-resolution data is a fundamental limiting factor. While convolutional neural networks (CNNs) with deep layers have achieved unprecedented improvements in object recognition applications that rely on fine image structures, they cannot be applied directly to medium-resolution data due to lack of such fine structures. In this paper, considering the spatial relation of a pixel to its neighborhood, we propose a new deep patch-based CNN system tailored for medium-resolution remote sensing data. The system is designed by incorporating distinctive characteristics of medium-resolution data; in particular, the system computes patch-based samples from multidimensional top of atmosphere reflectance data. With a test site from the Florida Everglades area (with a size of 771 square kilometers), the proposed new system has outperformed pixel-based neural network, pixel-based CNN and patch-based neural network by 24.36%, 24.23% and 11.52%, respectively, in overall classification accuracy. By combining the proposed deep CNN and the huge collection of medium-resolution remote sensing data, we believe that much more accurate land cover datasets can be produced over large areas.

1. Introduction

Land cover, which is the pattern of ecological resources and human activities dominating different areas of Earth's surface, is a critical type of data supporting many environmental science and land management applications at local, regional, and global scales (Foley et al., 2005; Meyer & Turner II, 1994). Given the importance of land cover information in global change and environmental sustainability research, there have been numerous efforts to produce land cover datasets at various scales (e.g., Bartholome & Belward, 2005; Fry et al., 2011; Gong et al., 2013; Homer et al., 2001; Jin et al., 2013; Vogelmann et al., 2001). Land cover patterns are observable and therefore can be mapped by ground surveys or remote sensing. While ground surveys are largely limited by logistical constraints, remote sensing, through the use of cameras and sensors mounted on aerospace-borne platforms, makes direct observations across large areas of the land surface, thus allowing land cover patterns to be mapped in a timely and cost-effective mode. Both visual interpretation and computer-based digital classification can be used to extract information on land cover from a variety of remote sensing data, such as aerial photographs, satellite imagery, thermal imagery, hyper-spectral imagery, radar and lidar datasets, which vary in spatial, spectral, radiometric, and temporal resolutions (Jensen, 2015). Digital pattern classification is generally preferred over visual interpretation for mapping land cover in large areas. With the availability of enormous amount of remote sensing data with improved spectral, spatial, and temporal resolutions, the real bottleneck is an accurate and effective image classification for remote sensing data. The remote sensing community has come a long way, starting with conventional methods such as maximum likelihood classifier (MLC) (Strahler, 1980) and clustering (Huang, 2002); it has now moved towards more advanced techniques such as decision trees (Xu, Watanachaturaporn, Varshney, & Arora, 2005), random forests (RF) (Pal, 2005), neural networks (NN) (Kavzoglu & Mather, 2003; Mas & Flores, 2008), support vector machines (SVM) (Gidudu, Hulley, & Marwala, 2007; Mountrakis, Im, & Ogole, 2011; Pal & Mather, 2005) and more recently convolutional neural networks (CNN) (Castelluccio, Poggi, Sansone, & Verdoliva, 2015; Romero, Gatta, & Camps-Valls, 2016) to classify remote sensing data.

* Corresponding author.
E-mail addresses: as13an@my.fsu.edu (A. Sharma), liux@cs.fsu.edu (X. Liu), xyang@fsu.edu (X. Yang), dishi@ku.edu (D. Shi).

http://dx.doi.org/10.1016/j.neunet.2017.07.017
0893-6080/

Very recently, deep networks have been demonstrated to achieve significant empirical improvements in fields like computer vision (He, Zhang, Ren, & Sun, 2015; Krizhevsky, Sutskever, & Hinton, 2012; Sermanet et al., 2013; Socher, Lin, Manning, & Ng, 2011) and natural language processing (Collobert & Weston, 2008; Hinton et al., 2012; Socher et al., 2011). For example, in computer vision, deep CNNs have surpassed human performance on the 1000-class ImageNet dataset (Russakovsky et al., 2015), which contains 1.2 million training images, 50,000 validation images, and 100,000 test images. Such effective techniques could have significant impacts on remote sensing image classification. Accordingly, the remote sensing community has also started to incorporate deep CNNs to image classification tasks (Castelluccio et al., 2015; Hu, Huang, Wei, Zhang, & Li, 2015; Nogueira, Penatti, & Santos, 2016; Penatti, Nogueira, & dos Santos, 2015; Romero et al., 2016; Zhao & Du, 2016). However, the majority of research using deep CNNs in the remote sensing community has been focusing on high-resolution images. Classification of these high-resolution images is similar to object recognition in computer vision, and remarkable improvements achieved by deep networks in object recognition have also been shown in these applications. However, in order to realize the full potential of deep networks for remote sensing, a large training dataset is needed (He et al., 2015; Krizhevsky et al., 2012) and it may be difficult to acquire a high-resolution dataset for a very large area. On the other hand, we have abundance of publicly available and free of charge remote sensing data at medium-resolution such as Landsat imagery (USGS, 2016a). Consequently, we have targeted medium-resolution Landsat imagery because of their overwhelming use as the primary data for global environmental change research. The data acquired by Landsat programs provide the longest continuous observations of Earth's surface from space. In particular, the Landsat system offers a rich archive of highly calibrated, multi-spectral data of global coverage that recently becomes available at no charge from the USGS EROS Data Center, which has become an invaluable resource for examining natural and anthropogenic changes on Earth's surface (USGS, 2016a; Yang, 2011). Therefore, a system that could combine the deep CNN and the huge collection of medium-resolution remote sensing data together will benefit many such applications. Furthermore, such a system would be able to provide more reliable and efficient classification of remote sensing data over a large area.

However, current deep neural networks, when applied to medium-resolution data using pixel based approaches, are not effective. Existing techniques to classify the medium-resolution remote sensing imagery are mostly pixel-based. For example, a 30 m resolution global land-cover dataset was produced using four different pixel-based classifiers, namely, MLC, decision trees, RF and SVM, with an overall accuracy of 64.9% (Gong et al., 2013). While the result can be useful for certain applications, it may not be sufficient to support applications that require more accurate land-cover information. A main objective of this research is to develop a deep network system for land cover classification of medium-resolution satellite imagery. Considering that a pixel in remote sensing imagery is spatially related to its neighborhood (Berberoglu, Lloyd, Atkinson, & Curran, 2000; Lloyd, Berberoglu, Curran, & Atkinson, 2004), we have developed a six-layer deep convolutional system with better accuracy which can be applied to medium-resolution imagery such as Landsat data (USGS, 2016a). With 30 m spatial resolution of Landsat images (non-thermal bands), the typical deep CNN architectures for object recognition may not be effective as they are normally used to detect small objects like houses, primarily based on shape, texture, and other fine structures. However, such features are not present in medium-resolution images. As a result, we have to make substantial changes to typical deep networks and create a unique deep network architecture to work with medium-resolution remote sensing data. Using a test site in a complex tropical area in Florida, the proposed new system has achieved 24.36%, 24.23% and 11.52% of improvement in overall classification accuracy over pixel-based neural network, pixel-based CNN and patch-based neural network, respectively (see Section 5.2 for details).

The remainder of this paper is organized as follows. In Section 2, we discuss about remote sensing imagery briefly. We describe convolutional neural networks customized for remote sensing applications in Section 3. In Section 4, a new patch-based CNN we develop for classifying medium-resolution remote sensing data is presented. The experimental results from the proposed system and the comparisons with pixel-based conventional neural network, pixel-based CNN and multidimensional patch-based neural network with the same spatial context are given in Section 5. Finally, Section 6 summarizes the major findings and discusses some issues for further research.

2. Overview of remote sensing imagery

The purpose of this brief section is to provide essential information about medium-resolution remote sensing imagery related to the proposed deep network architecture; interested readers are referred to Jensen (2015) and Lillesand, Kiefer, and Chipman (2008) for further information. As mentioned earlier, remote sensing provides an effective and scalable way for mapping land cover patterns. In general, remote sensing imagery is the collection of spectral and thermal bands acquired using different sensor systems (Jensen, 2015); each band represents a specific wavelength region of the Electromagnetic Spectrum (EMS). This paper uses Landsat 8 medium-resolution remote sensing imagery. Landsat 8 is the latest edition to the Landsat program and the data products of Landsat 8 represent 9 spectral bands acquired by the Operational Land Imager (OLI) (covering the visible, near infrared and short-wave infrared bands) and 2 additional thermal bands acquired by the Thermal Infrared Sensor (TIRS) using quantized and calibrated scaled Digital Numbers (DN). Each Landsat 8 data product consists of 8 OLI bands with a spatial resolution of 30 m, one OLI band (panchromatic) with 15 m spatial resolution, 2 thermal bands with 100 m resolution and one product metadata file (MTL file). In general, different sensors and platforms may cover different wavelength regions as bands. From an image analysis point of view, the resulting images corresponding to different wavelength regions of the EMS are saved digitally and available as multiband images.

In remote sensing, classification schemes are developed to classify remotely sensed data successfully into land cover or use information with different levels of classification details (e.g., USGS Land-Use/Land-Cover and LBCS) (Jensen, 2015). The required level of detail decides the desired spatial resolution of the remote sensory data; spatial resolution is the measure of the smallest angular or linear separation between two objects to which sensor is sensitive (e.g., Landsat 8 imagery is of 30 m spatial resolution and Digital globe's QuickBird has 2.44 m spatial resolution). Medium-resolution remote sensing imagery ranges between 10 m and 60 m (e.g., ASTER, EO-1, Landsat and Sentinel) and is appropriate to extract region and biome level of details (Jensen, 2015). In case of regional level classification, the spatial context plays a major role and provides inherent information about the point of interest (e.g. forest, wetland and cropland); however, the existing classifiers for medium-resolution images are mainly pixel based. In this paper, we demonstrate the importance of spatial context for classifying the medium-resolution images by using patch-based samples.

3. Convolutional neural networks

In conventional multilayer neural networks, input is given as a single vector and is transformed over a series of hidden layers to reach the output. The hidden layers comprise of a chosen number of hidden units (neurons) and each neuron in a hidden layer is fully connected to all the neurons present in the previous layer independently. Several previous studies on land cover classification using multilayer neural networks have achieved satisfactory results (Kavzoglu & Mather, 2003; Mas & Flores, 2008; Shupe & Marsh, 2004). However, the standard multilayer neural networks are limited when dealing with multidimensional images. Firstly, an enormous number of parameters are needed for multidimensional inputs because all input elements are converted into a single vector. For example, an image of size $X \times Y$ with Z bands is equivalent to $X \times Y \times Z$ inputs, which requires to calculate $X \times Y \times Z$ weights in the first hidden layer. With additional hidden layers, these parameters add up quickly. Secondly, spatial contexts between pixels are not considered explicitly in the multilayer neural network; all pixels are considered independently without considering their spatial locality in the image. Convolutional Neural Networks (CNNs) are a variant of multilayer neural networks, inspired by the animal's visual cortex (Hubel & Wiesel, 1968). CNNs handle images as a multidimensional input, instead as a single vector and consider the spatial contexts of image pixels explicitly. CNNs are characterized by several unique properties that will be discussed below.

3.1. Local connectivity

In CNN, a neuron in a hidden layer is connected only to a subregion (called receptive field) of the input unlike the conventional multilayer neural network where neuron is connected to all the neurons in the previous layer; therefore, fewer parameters are required and less computation is needed consequently. The structure of neurons in the hidden layer is defined by the number of channels present in the receptive field. Note that the local receptive field of an neuron in the second layer covers a larger area in terms of input than a related neuron in the first layer. By having more layers than conventional neural networks, CNN can also model long range dependencies but more efficiently (Fukushima & Miyake, 1982; Hubel & Wiesel, 1968).

3.2. Parameter sharing

Parameter sharing is also an important property of CNN, where all the neurons belonging to a particular feature map share the same weighted connections and these neurons cover all the used receptive fields. The weighted connections can also be seen as a filter or kernel. With parameters shared by all the neurons belonging to a feature map, this property reduces the number of the parameters substantially. A feature map using an unique filter extracts the same feature at all the receptive fields. Different feature maps together create a multidimensional matrix which generates inputs for the next layer (Fukushima & Miyake, 1982).

3.3. Pooling/subsampling

Pooling/subsampling layer is a combination of two operations. Pooling is performed on non overlapping patches present in the feature map, in which a pool of neighboring neurons is replaced by a single neuron. In subsampling, those pooling generated neurons replace the pool of neighboring neurons in the next hidden layer. It reduces the number of neurons using pooling. Pooling can be done using different functions, such as the maximum, minimum, and average of the involved pixels (Boureau, Ponce, & LeCun, 2010).

3.4. Depth

Depth of the CNNs, i.e., the number of layers, is also very important. Recent CNNs advocate the use of deep architectures (He et al., 2015; Krizhevsky et al., 2012). Theoretically, the advantages of having a deep architecture over a conventional shallow architecture are not yet fully understood (Bengio, LeCun, et al., 2007). However, the deep architecture is preferred to solve complex AI problems like image analysis and natural language processing (Bengio, 2009), where a problem is decomposed into multilevels of computation and representation. In order to represent same functions, studies also show that the growth of the number of units in deep networks is linear, compared to exponential growth in case of shallow networks (Delalleau & Bengio, 2011). The proposed architecture supports the same idea, where classification accuracy of the system decreases even if a single convolution layer is removed.

These features together enable CNN to model complex relationships among input elements efficiently by creating different combinations along paths through the layers. Compared to conventional neural networks, CNNs avoid the need of exponentially many neurons in hidden layers. With effective training algorithms developed in recent years, they have led to unprecedented performance in many applications (e.g., Collobert & Weston, 2008; He et al., 2015; Hinton et al., 2012; Krizhevsky et al., 2012; Sermanet et al., 2013; Socher et al., 2011).

4. A patch-based CNN system for remote sensing image classification

The proposed system is adapted for medium-resolution remote sensing imagery. While the proposed system is generic and should work for all the medium-resolution multidimensional data, we have tested the new system on Landsat images. Below, we define the features used and the architecture adopted.

4.1. Multidimensional data

Landsat 8 imagery is categorized as a medium-resolution imagery with 30 m spatial resolution for the non-thermal bands. With this resolution, we cannot use the typical object recognition deep CNN architecture where it is used to detect small objects like houses, primarily, based on their shapes and fine structures. In order to deal with this limitation, we have calculated the top-of-atmosphere (TOA) reflectance values associated with the pixels from the scaled Digital Numbers (DN) belonging to all the OLI bands (except the panchromatic band). TOA reflectance values can be obtained by rescaling and correcting the default 16-bit unsigned integer format DN values using radiometric (reflectance) rescaling coefficients and sun angle provided in the MTL file present with Landsat 8 product (USGS, 2016b). A Landsat image is then converted into a multidimensional TOA reflectance vector of size $X \times Y \times Z$ where X specifies the width, Y the height, and Z the number of channels. In the proposed system, we have calculated the TOA reflectance at all the locations for the eight OLI bands independently; as a result, we have eight different TOA reflectance 2D matrices of size $X \times Y$. Later, we club these eight TOA reflectance 2D matrices into one eight-channel multidimensional TOA reflectance vector, where each location at this multidimensional vector represents a vector of TOA reflectances of length eight. This representation is used as input instead of the multi-spectral Landsat imagery. Values of X, Y, and Z vary with the medium-resolution imagery source but the structure of multidimensional data remains the same.

4.2. Patch-based samples

CNN requires image-like multidimensional input instead of a single vector. Therefore, we have to extract multidimensional samples instead of pixel-based single vector samples. To do that, samples are extracted as patches with size $5 \times 5 \times 8$ out of multidimensional data and labeled using the center pixel of each patch. These patches have multidimensional image-like structures of TOA reflectance data which are different from the usual segments representing objects or groups of objects in object based image analysis for high-resolution remote sensing data (Blaschke, 2010; Blaschke et al., 2014). Optimal patch size can vary with the medium-resolution imagery source; however, we have found that the window size of $5 \times 5 \times 8$ size is able to capture spatially local correlation of a center pixel to the surrounding pixels and limit heterogeneous pixels in case of Landsat 8. Patches are extracted for all possible locations in the data, which overlap with the neighboring patches. In order to extract maximal samples, the stride value is set to one while extracting patches and all the valid locations (except the boundary and the cloud/shadow surrounded locations) are used to extract samples.

4.3. Cloud/shadow mask

In order to deal with cloud/shadow pixels, a cloud/shadow mask is generated for a Landsat image using the Fmask Algorithm (Zhu & Woodcock, 2012). The mask is then used to locate all the cloud/shadow pixels present in the imagery. TOA reflectance vector of the cloud/shadow location is set to zero. If a patch contains any cloud/shadow pixels, it is marked as unused.

4.4. Training sample selection

While extracting training samples out of all the available patches, we impose two additional constraints. First, 60% or more of pixels present in a patch should belong to the same class as the center pixel. Second, there should not be any cloud/shadow pixel present in the patch. Samples are extracted for the locations in the data which satisfy the constraints.

4.5. Stride/Padding/Filters/Depth

In the proposed architecture, the size of input samples is only $5 \times 5 \times 8$. Therefore, all the convolutions have stride value one and are zero padded to make the size of the output same as the input. The number of filters is doubled with each subsequent convolutional layer to increase the number of feature maps in the hidden layers.

The proposed CNN consists of six layers and an additional softmax layer. The first five hidden layers are convolutional and the last hidden layer is fully connected. We have achieved best results with five convolutional layers. Classification accuracy of the system decreases even if a single convolution layer is removed. However, including more than five convolution layers does not show any significant improvement in the classification accuracy.

4.6. Architecture

Fig. 1 illustrates the overall architecture of the proposed CNN system. We have not used any pooling/subsampling layer because the size of our samples is only $5 \times 5 \times 8$. The softmax layer generates a probability distribution over the eight classes, using the output from the fully connected layer as its input. To implement this network, we have used Google's tensorflow (Abadi et al., 2015) (an open source software library for machine intelligence) and Quadro K5200 GPU. The proposed network minimizes the cross entropy

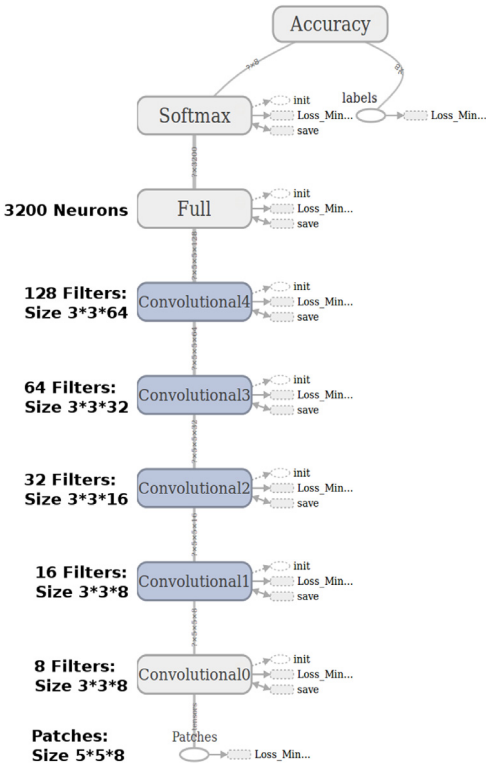

Fig. 1. Architecture of the proposed convolutional neural network system.

using the ADAM optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001. The ADAM optimizer is a first-order gradient-based optimization algorithm of stochastic objective functions; stochastic gradient descent proves to be very efficient and effective optimization method in recent deep learning networks (Deng et al., 2013; Hinton et al., 2012; Hinton & Salakhutdinov, 2006; Krizhevsky et al., 2012). A rectified linear unit (ReLU) activation function (Nair & Hinton, 2010) is applied on the outputs generated from all the convolutional and fully connected layers.

Convolutional layer 0 takes $5 \times 5 \times 8$ input patch and filters them using eight filters of size $3 \times 3 \times 8$. All the other convolutional layers take the output of the previous hidden layer as its input. The number of filters is doubled with each subsequent convolutional layer; convolutional layer 1 has 16 filters, convolutional layer 2 has 32 filters, and so on. With samples of size $5 \times 5 \times 8$, we are able to extract all the relevant information from the input data by doubling the number of filters with each subsequent convolutional layer. The size of all the filters belonging to one convolutional layer is the same. Filter size at convolutional layer 1 is same as convolution layer 0. Convolution layer 2 filter size is $3 \times 3 \times 16$, convolution layer 3 filter size is $3 \times 3 \times 32$ and the last convolutional layer has 128 filters of size $3 \times 3 \times 64$. The fully connected layer has 3200 (given by $5 \times 5 \times 128$) neurons.

Fig. 2 shows the evolution of loss minimization (left) and the classification accuracy (right) on the training data. The initial loss value at step 0 is 11.8 in the proposed architecture; in order to improve readability, the plot starts from step 300. We have also produced classification results for the whole area at regular intervals to visualize the improvements in the classification accuracy of the proposed system over time. Fig. 3 shows the improvements in classification results over time. It shows classification results after certain number of iterations and the reference map. Using the reference map as ground truth, the accuracy, computed as the percentage of the pixels with the same label as the reference map

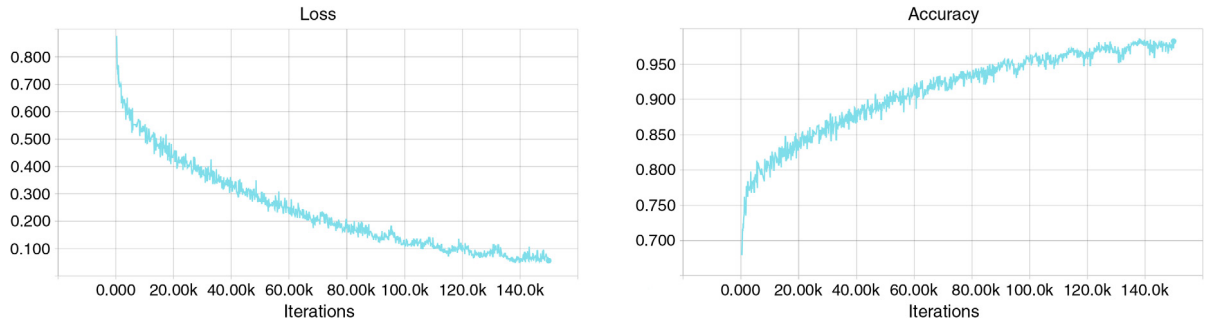

Fig. 2. Loss (left) and accuracy (right) plots for the training data.

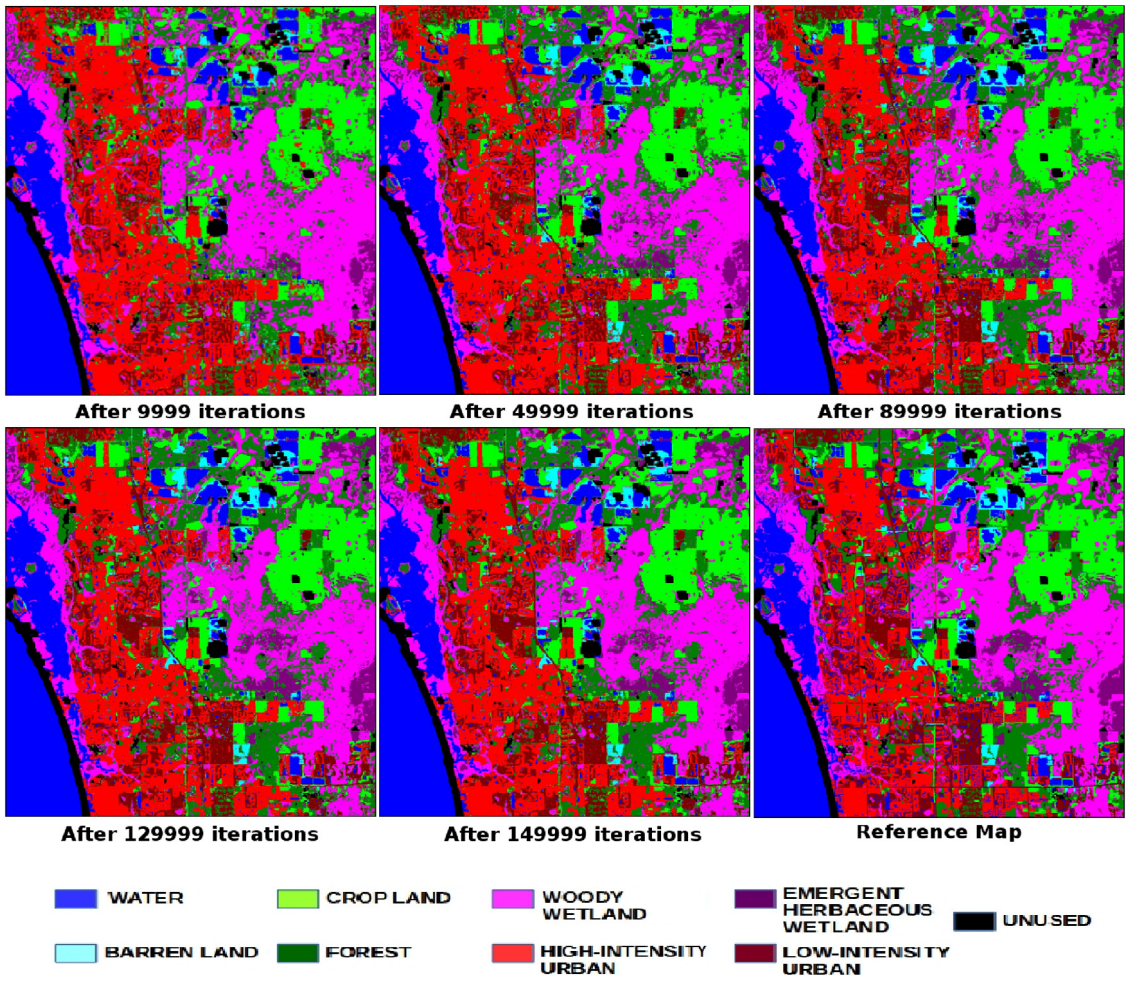

Fig. 3. Illustration of improvement in classification results with number of iterations. Compared to reference map, after certain number of iterations, the accuracy is as follows: (a) 9999 iterations: 72.93%, (c) 49 999 iterations: 79.12%, (d) 89 999 iterations: 83.06%, (e) 129 999 iterations: 84.67%, (f) 149 999 iterations: 85.60%.

is 72.93%, 79.12%, 83.06%, 84.67% and 85.60% after 9999, 49 999, 89 999, 129 999 and 149 999 iterations, respectively.

5. Experimental results and comparisons

5.1. Test site

We choose to implement the proposed system on a test site within the Florida Everglades ecosystem; this ecosystem has attracted international attention for the ecological uniqueness and fragility. It is comprised of a wide variety of sub-ecosystems such as freshwater marshes, tropical hardwood hammocks, cypress swamps, and mangrove swamps (Davis & Ogden, 1994). Such diverse ecological types make the Everglades an ideal site to test the reliability and robustness of this new system. The Landsat8 image in our study was acquired on February 10, 2014 (Path 16; Row 42); we extract a subset from this image as our test site with a size equivalent to 771 square kilometers. Each pixel in this subset (with 864 $\times$ 991 pixels) has 30 m $\times$ 30 m spatial resolution.

In order to create a reference map for our research area, we obtained ancillary data from the Florida Cooperative Land Cover Map first and performed correction by comparing it with GPS-guided field observations and the high-resolution images from Google Earth. We have used this reference map to generate training samples and perform accuracy assessments (Lo & Watson, 1998). Using the ancillary data, we adopt a mixed Anderson Level 1/2

Table 1
Land cover classification scheme and training sample size.

No.	Class name	Description	Training samples
1	High intensity urban	Commercial, industrial, institutional constructions with large roofs. Large open spaces and large transportation facilities. Residential areas with impervious surfaces more than half of the total cover.	112 582
2	Low intensity urban	Residential areas with impervious surfaces less than half of the total cover. Smaller urban service buildings, such as detached stores and restaurants, and state highways.	48 744
3	Barren land	Urban areas with low percentages of constructed materials, vegetation, and low level of impervious surfaces, including bare soil lands, beaches.	10 228
4	Forest	Herbaceous cover, trees, trees remain green throughout the year, some wetland evergreen forests included.	81 740
5	Cropland	Crops and pastures with vegetation coverage mixed with bushes, small amount fallow land.	79 628
6	Woody wetland	Cypress/tupelo, strand swamp, other coniferous wetland, mixed wetland hardwoods, mangrove swamp.	151 533
7	Emergent herbaceous wetland	Freshwater non-forested wetland, prairies/bogs, freshwater marshes, wet prairies, saltwater marsh.	30 389
8	Water	Streams, canals, lakes, ponds, bays.	108 594

land-use/land-cover classification scheme (Anderson, 1976) with eight classes (see Table 1). Based on our training sample extraction constraints, we are able to generate training samples of size $5 \times 5 \times 8$ for the eight classes; the details are shown in Table 1.

5.2. Experimental results/comparisons

In this section, we present the experimental results of the proposed system and compare them from that a pixel-based conventional neural network, a pixel-based CNN and a multidimensional patch-based neural network. By comparing pixels directly in each of the classification maps from the four different networks for the whole area to the reference map, pixel-based conventional neural network achieves accuracy of 62.34%, pixel-based CNN 63.01%, patch-based neural network 73.17%, and the proposed system 85.60%. We have also tested the classification accuracies of SVM (one-vs-one) and RF classifiers for the whole area using the above approach. Python's scikit-learn package[1] is used to implement these two classifiers. In case of SVM and RF, the achieved accuracies are 61.86% and 75.22%, respectively, which are still much lower than the proposed system. However, in order to perform quantitative evaluation of the classification results generated by four different neural networks to determine overall and individual category classification accuracy, we have done accuracy assessment using the method described by Congalton (Congalton, 1991). Specifically, error matrix is generated using the weighted random stratified sampling and then the overall accuracy (OA), Overall kappa (KAPPA), Producer's accuracy (PA), User's accuracy (UA) and conditional kappa are calculated based on the error matrix (Jensen, 2015).

The proposed system and multidimensional patch-based system use the same samples, which are in the form of multidimensional TOA reflectance patches of size $5 \times 5 \times 8$. However, pixel-based conventional and CNN systems use only the center pixel vector of each patch and this pixel vector is of length 8, containing the TOA reflectance values of 8 OLI bands at that location. Therefore, the number of training samples is the same in all the four networks as mentioned in Table 1. The proposed system achieves 89.26% in the overall accuracy and 0.874 in Kappa index. For the individual categories, this new system achieves PA and UA values more than 80% for all classes and the mean of conditional kappa values belonging to 8 different classes (Mean-Kappa) is 0.88 with minimum 0.78 conditional kappa index. In some classes, the system achieves significant improvements. For example cropland and woody wetland have 91.82%, 94.39%, 0.94 and 89.9%, 93.97%, 0.92 as PA, UA and conditional kappa index values, respectively. The proposed system is able to achieve good results for several spectrally and spatially complex classes, such as the low intensity urban and cropland. Table 2 shows the complete error matrix of system with calculated OA, KAPPA of the overall system and PA, UA, conditional kappa for all individual classes separately.

The new patch-based CNN system achieves better results than the standard pixel-based neural network, the pixel-based CNN and the patch-based neural network. Comparative results are summarized in Table 3. Fig. 4 shows the comparison to reference map from the classification results of four different systems. The proposed system gets 89.26% OA, 0.87 KAPPA, 0.88 Mean-Kappa, which achieves 11.52%, 0.13, 0.16 and 24.23%, 0.28, 0.16 and 24.36%, 0.29, 0.31 improvements over patch-based neural network, pixel-based CNN and standard pixel-based neural network, respectively. Both pixel-based conventional and CNN systems achieve almost similar results; so, there is no significant improvement by using CNN over conventional neural network on pixel-based system. Table 3 also shows that there are significant enhancements not only in the overall but also in all the individual categories by the proposed system. Table 3 shows that the proposed system achieves a minimum 0.23 and 0.26 increase in conditional kappa values for all the classes except water when comparing with the pixel-based neural network and pixel-based CNN, respectively; where five classes have more than 0.33 improvements in the conditional kappa values when comparing with both of the pixel-based systems. In comparison to the patch-based neural network without CNN, the proposed system achieves a minimum 0.13 increase in conditional kappa values for six classes with maximum 0.25. The proposed system shows substantial improvements in the conditional kappa results for hard-to-classify classes also. For example, in case of low intensity urban and cropland, there are 0.14, 0.29, 0.33 and 0.13, 0.34, 0.37 improvement over patch-based neural network, pixel-based CNN and standard pixel-based neural network, respectively.

Several previous studies have suggested single hidden layer neural networks perform better for classification of remote sensing images (Kanellopoulos & Wilkinson, 1997; Shupe & Marsh, 2004). Therefore, we have used only one fully connected hidden layer between input and softmax (outer) layer for both pixel-based neural network and multidimensional patch-based neural network. In both cases, the hidden layer consists of 200 (given by $5 \times 5 \times 8$) neurons. In order to implement pixel-based CNN, we have added an addition convolutional layer with 8 filters of size $1 \times 1 \times 8$. Similar to the new patch-based CNN system, we have used Google's tensorflow (Abadi et al., 2015) and Quadro K5200 GPU to implement these three networks also. As shown in Table 3 and Fig. 4, the patch-based neural network improves significantly over both pixel-based systems in OA, KAPPA, Mean-Kappa of the overall

[1] Obtained from http://scikit-learn.org/stable/.

Table 2
Error matrix using the new deep patch-based CNN system.

Classified data	Reference data									Accuracy and conditional kappa		
	High intensity urban	Low intensity urban	Barren land	Forest	Crop land	Woody wetland	Emergent herbaceous wetland	Water	Row total	Producer's accuracy (PA%)	User's accuracy (UA%)	Conditional kappa
High intensity urban	**149**	11	1	2	4	3	1	8	179	95.51	83.24	0.80
Low intensity urban	5	**75**	0	3	2	0	2	2	89	82.42	84.27	0.83
Barren land	0	1	**42**	1	0	1	2	3	50	91.30	84	0.83
Forest	1	1	2	**107**	2	13	3	4	133	88.43	80.45	0.78
Cropland	1	2	0	0	**101**	0	2	1	107	91.82	94.39	0.94
Woody wetland	0	1	0	8	1	**187**	1	1	199	89.90	93.97	0.92
Emergent herbaceous wetland	0	0	0	0	0	4	**46**	0	50	80.70	92	0.91
Water	0	0	1	0	0	0	0	**132**	133	87.42	99.25	0.99
Column total	156	91	46	121	110	208	57	151	**940**			

Overall accuracy(OA): **89.26%**; Overall kappa(KAPPA): **0.874**

Table 3
Summary of the accuracy assessment for the classification results produced by the new deep patch-based CNN system (Deep CNN), pixel-based neural network system (Pixel NN), pixel-based CNN (Pixel CNN) and multidimensional patch-based network system (Patch NN).

Land cover class	Conditional Kappa					
	Pixel NN	Pixel CNN	Patch NN	Deep CNN	Mean	Standard deviation
High intensity urban	0.57	0.54	0.72	0.80	0.66	0.12
Low intensity urban	0.50	0.54	0.69	0.83	0.64	0.15
Barren land	0.56	0.44	0.64	0.83	0.62	0.16
Forest	0.41	0.44	0.56	0.78	0.55	0.17
Cropland	0.57	0.6	0.81	0.94	0.73	0.18
Woody wetland	0.52	0.57	0.76	0.92	0.69	0.18
Emergent herbaceous wetland	0.57	0.46	0.66	0.91	0.65	0.19
Water	0.93	0.94	0.96	0.99	0.96	0.03
Mean-kappa	0.58	0.57	0.72	0.88		
Standard deviation	0.15	0.16	0.12	0.08		
Overall accuracy(%)	64.9	65.03	77.74	89.26		
Overall kappa	0.58	0.59	0.74	0.87		

Table 4
Error matrix using pixel-based neural network.

Classified data	Reference data									Accuracy and conditional kappa		
	High intensity urban	Low intensity urban	Barren land	Forest	Crop land	Woody Wetland	Emergent herbaceous wetland	Water	Row total	Producer's accuracy (PA%)	User's accuracy (UA%)	Conditional kappa
High intensity urban	**134**	25	7	14	7	4	3	16	210	81.21	63.81	0.57
Low intensity urban	8	**27**	0	5	5	2	0	3	50	33.33	54	0.50
Barren land	1	2	**29**	6	3	3	3	3	50	64.44	58	0.56
Forest	3	1	3	**49**	6	25	12	1	100	36.03	49	0.41
Cropland	14	10	0	16	**77**	6	1	1	125	74.04	61.60	0.57
Woody wetland	4	14	3	40	5	**168**	28	6	268	76.02	62.69	0.52
Emergent herbaceous wetland	0	1	1	5	1	11	**30**	1	50	38.46	60	0.57
Water	1	1	2	1	0	2	1	**128**	136	80.50	94.12	0.93
Column total	165	81	45	136	104	221	78	159	**989**			

Overall accuracy (OA): **64.9%**; Overall kappa (KAPPA): **0.584**

Table 5
Error matrix using pixel-based CNN.

Classified data	Reference data									Accuracy and conditional kappa		
	High intensity urban	Low intensity urban	Barren land	Forest	Crop land	Woody Wetland	Emergent herbaceous wetland	Water	Row total	Producer's accuracy (PA%)	User's accuracy (UA%)	Conditional kappa
High intensity urban	**125**	33	5	10	17	2	2	10	204	78.61	61.27	0.54
Low intensity urban	8	**29**	0	3	1	4	2	3	50	31.18	58	0.54
Barren Land	4	2	**23**	10	4	3	3	1	50	71.87	46	0.44
Forest	4	5	0	**53**	9	18	12	1	102	38.41	51.96	0.44
Cropland	8	11	0	14	**85**	3	6	5	132	69.11	64.39	0.60
Woody wetland	7	12	1	41	4	**171**	16	6	258	80.66	66.28	0.57
Emergent herbaceous wetland	1	1	0	7	3	9	**25**	4	50	37.88	50	0.46
Water	2	0	3	0	0	2	0	**125**	132	80.65	94.70	0.94
Column total	159	93	32	138	123	212	66	155	**978**			

Overall accuracy (OA): **65.03%**; Overall kappa (KAPPA): **0.587**

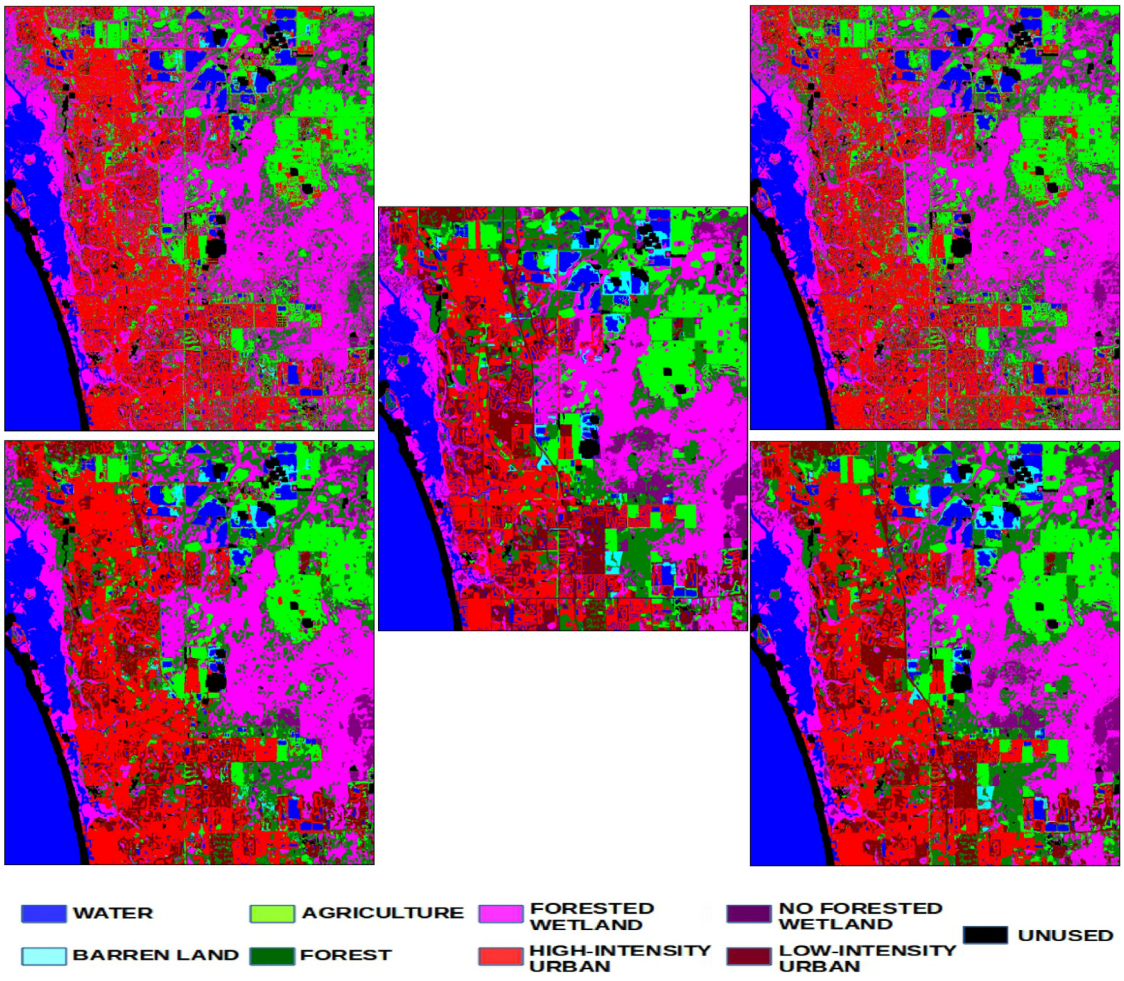

Fig. 4. Classification results for the study site from four different methods: (a) Upper left: Pixel-Based Neural Network, (b) Upper Right: Pixel-Based CNN, (c) Lower right: Patch-Based Neural Network and (d) Lower right: New Deep Patch-Based CNN, (e) Center: Reference Map.

Table 6
Error matrix using patch-based neural network.

Classified data	Reference data									Accuracy and conditional kappa		
	High intensity urban	Low intensity urban	Barren land	Forest	Crop land	Woody Wetland	Emergent herbaceous wetland	Water	Row total	Producer's accuracy (PA%)	User's accuracy (UA%)	Conditional kappa
High intensity urban	**145**	18	4	3	9	2	0	8	189	88.41	76.72	0.72
Low intensity urban	7	**65**	1	6	5	0	3	3	90	71.43	72.22	0.69
Barren land	4	2	**33**	2	4	0	3	2	50	78.57	66	0.64
Forest	4	2	1	**87**	7	32	6	1	140	68.50	62.14	0.56
Cropland	2	2	1	5	**85**	3	3	1	102	76.58	83.33	0.81
Woody wetland	1	2	1	17	0	**173**	13	5	212	79.36	81.60	0.76
Emergent herbaceous wetland	0	0	0	7	1	7	**34**	1	50	53.97	68	0.66
Water	1	0	1	0	0	1	1	**129**	133	86	96.99	0.96
Column Total	164	91	42	127	111	218	63	150	**966**			

Overall accuracy (OA): **77.74%**; Overall kappa (KAPPA): **0.738**

system and PA, UA, conditional kappa for all individual classes. In the patch-based neural network, there are 12.84% and 12.71% improvement in OA, 0.16 and 0.15 improvement in KAPPA, and 0.14 and 0.15 improvement in Mean-Kappa, comparing to pixel-based neural network and pixel-based CNN, respectively. The same weighted stratified sampling is used for accuracy assessments in these networks too. The details of the error matrices, OA, KAPPA, PA, UA and conditional kappa are given in Table 4 for pixel-based neural network, Table 5 for pixel-based CNN and Table 6 for patch-based network, respectively. The substantial improvements will lead to more accurate land cover data that are essential for many applications (e.g., agriculture monitoring, energy development and resource exploration).

6. Conclusion and future work

In this paper, we have proposed a new patch-based CNN system tailored for medium-resolution remote sensing imagery classification. The proposed system uses new features to adapt it for remote sensing data classification. Specifically for Landsat data, we have computed patch-based samples from multidimensional TOA

reflectance data. The proposed system is compared to the pixel-based conventional neural network, the pixel-based CNN and the patch-based neural network. The classification results show that the proposed system achieves significant improvements in both the overall and categorical classification accuracies.

There are several changes that could lead to further improvements. For example, classification accuracy could improve further by creating hierarchical structure classification on the top of the proposed system using different sizes of patches for the same center pixel. We believe that the use of multi-temporal remote sensing data over the same area could improve the performance even more. These and other parameter choices are being investigated further.

In order to apply this new CNN system on the global level, the efficiency of the convolution operations can be improved substantially by incrementally updating the convolutions (Sermanet et al., 2013). As the softmax layer outputs a probability distribution, contextual constraints can be imposed on the class labels to improve the boundary accuracy. We believe that the proposed system can lead to much more accurate global maps.

Acknowledgments

The authors would like to thank the anonymous reviewers whose comments and suggestions have improved the presentation of this paper significantly.

References

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, & Citro, Craig et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Anderson, James Richard (1976). *A land use and land cover classification system for use with remote sensor data, Vol. 964*. US Government Printing Office.

Bartholome, E., & Belward, A. S. (2005). Glc2000: a new approach to global land cover mapping from earth observation data. *International Journal of Remote Sensing*, *26*(9).

Bengio, Yoshua (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127.

Bengio, Yoshua, LeCun, Yann, et al. (2007). Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, *34*(5).

Berberoglu, S., Lloyd, Christopher D., Atkinson, P. M., & Curran, Paul J. (2000). The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. *Computers & Geosciences*, *26*(4), 385–396.

Blaschke, Thomas (2010). Object based image analysis for remote sensing, *International Journal of Photogrammetry and Remote Sensing*, *65*(01).

Blaschke, Thomas, Hay, Geoffrey J., Kelly, Maggi, Lang, Stefan, Hofmann, Peter, Addink, Elisabeth, et al. (2014). Geographic object-based image analysis - towards a new paradigm. *International Journal of Photogrammetry and Remote Sensing*, *87*(0).

Boureau, Y-Lan, Ponce, Jean, & LeCun, Yann (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning, ICML-10* (pp. 111–118).

Castelluccio, Marco, Poggi, Giovanni, Sansone, Carlo, & Verdoliva, Luisa (2015). Land use classification in remote sensing images by convolutional neural networks, ArXiv Preprint arXiv:1508.00092.

Collobert, Ronan, & Weston, Jason (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167). ACM.

Congalton, Russell G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46.

Davis, Steve, & Ogden, John C. (1994). *Everglades: the ecosystem and its restoration*. CRC Press.

Delalleau, Olivier, & Bengio, Yoshua (2011). Shallow vs. deep sum-product networks. In *Advances in neural information processing systems* (pp. 666–674).

Deng, Li, Li, Jinyu, Huang, Jui-Ting, Yao, Kaisheng, Yu, Dong, Seide, Frank, et al. (2013). Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8604–8608). IEEE.

Foley, Jonathan A., DeFries, Ruth, Asner, Gregory P., Barford, Carol, Bonan, Gordon, Carpenter, Stephen R., et al. (2005). Global consequences of land use. *Science*, *309*(5734).

Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., Yang, L., et al. (2011). Completion of the 2006 national land cover database for the conterminous united states. *Photogrammetric Engineering and Remote Sensing*, *77*.

Fukushima, Kunihiko, & Miyake, Sei (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Springer.

Gidudu, Anthony, Hulley, Greg, & Marwala, Tshilidzi (2007). Classification of images using support vector machines, CoRR (arXiv:0709-3967v1).

Gong, Peng, Wang, Jie, Yu, Le, Zhao, Yongchao, Zhao, Yuanyuan, Liang, Lu, et al. (2013). Finer resolution observation and monitoring of global land cover: first mapping results with landsat tm and etm+ data. *International Journal of Remote Sensing*, *34*(7).

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2015). Deep Residual Learning for Image Recognition, ArXiv Preprint arXiv:1512.03385.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel rahman, Jaitly, Navdeep, et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hinton, Geoffrey E., & Salakhutdinov, Ruslan R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., et al. (2001). Completion of the 2001 national land cover database for the conterminous united states. *Photogrammetric Engineering and Remote Sensing*, *73*.

Hu, Wei, Huang, Yangyu, Wei, Li, Zhang, Fan, & Li, Hengchao (2015). Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, *2015*.

Huang, Kal-Y. I. (2002). A synergistic automatic clustering technique (SYNERACT) for multispectral image analysis. *Photogrammetric Engineering and Remote Sensing*, *68*(1).

Hubel, David H., & Wiesel, Torsten N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Jensen, John R. (2015). *Introductory digital image processing: a remote sensing perspective*. (4th ed.). Pearson Education, Inc..

Jin, Suming, Yang, Limin, Danielson, Patrick, Homer, Collin, Fry, Joyce, & Xian, George (2013). A comprehensive change detection method for updating the national land cover database to circa 2011. *Remote Sensing of Environment*, *132*(0).

Kanellopoulos, I., & Wilkinson, G. G. (1997). Strategies and best practice for neural network image classification. *International Journal of Remote Sensing*, *18*(4), 711–725.

Kavzoglu, T., & Mather, P. M. (2003). The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing*, *24*(23).

Kingma, Diederik, & Ba, Jimmy (2014). Adam: A method for stochastic optimization, ArXiv Preprint arXiv:1412.6980.

Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems, Vol. 25*. Curran Associates, Inc..

Lillesand, Thomas M., Kiefer, Ralph W., & Chipman, Jonathan W. (2008). *Remote sensing and image interpretation*. (6th ed.). John Wiley and Sons, Inc..

Lloyd, Christopher D., Berberoglu, S., Curran, Paul J., & Atkinson, Peter M. (2004). A comparison of texture measures for the per-field classification of Mediterranean land cover. *International Journal of Remote Sensing*, *25*(19), 3943–3965.

Lo, C. P., & Watson, Lee J. (1998). The influence of geographic sampling methods on vegetation map accuracy evaluation in a swampy environment. *Photogrammetric Engineering and Remote Sensing*, *64*(12), 1189–1200.

Mas, J. F., & Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, *29*(3).

Meyer, William B., & Turner II, B. L. (1994). *Changes in land use and land cover: a global perspective, Vol. 4*. Cambridge University Press.

Mountrakis, Giorgos, Im, Jungho, & Ogole, Caesar (2011). Support vector machines in remote sensing: A review. *International Journal of Photogrammetry and Remote Sensing*, *66*.

Nair, Vinod, & Hinton, Geoffrey E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning, ICML-10* (pp. 807–814).

Nogueira, Keiller, Penatti, Otávio A. B., & dos Santos, Jefersson A. (2016). Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification, ArXiv Preprint arXiv:1602.01517.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1).

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*(5), 1007–1011.

Penatti, Otávio A. B., Nogueira, Keiller, & dos Santos, Jefersson A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44–51).

Romero, Adriana, Gatta, Carlo, & Camps-Valls, Gustau (2016). Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1349–1362.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sermanet, Pierre, Eigen, David, Zhang, Xiang, Mathieu, Michaël, Fergus, Rob, & LeCun, Yann (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks, ArXiv Preprint arXiv:1312.6229.

Shupe, Scott M., & Marsh, Stuart E. (2004). Cover-and density-based vegetation classifications of the Sonoran Desert using Landsat TM and ERS-1 SAR imagery. *Remote Sensing of Environment*, *93*(1), 131–149.

Socher, Richard, Lin, Cliff C., Manning, Chris, & Ng, Andrew Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning, ICML-11* (pp. 129–136).

Strahler, Alan H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment*, *10*(2).

USGS. (2016a). Landsat Data Access, http://landsat.usgs.gov/Landsat_Search_and_Download.php [Online; accessed 11.08.16].

USGS. (2016b). Using the USGS Landsat 8 Product, http://landsat.usgs.gov/Landsat8_Using_Product.php [Online; accessed 11.08.16].

Vogelmann, James E., Howard, Stephen M., Yang, Limin, Larson, Charles R., Wylie, Bruce K., & Van Driel, Nicholas J. (2001). Completion of the 1990s national land cover data set for the conterminous united states from landsat thematic mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing*, *67*(6).

Xu, Min, Watanachaturaporn, Pakorn, Varshney, Pramod K., & Arora, Manoj K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*(3).

Yang, Xiaojun (2011). Use of Archival Landsat Imagery to Monitor Urban Spatial Growth. *Urban Remote Sensing: Monitoring, Synthesis and Modeling in the Urban Environment*, 15–33.

Zhao, Wenzhi, & Du, Shihong (2016). Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4544–4554.

Zhu, Zhe, & Woodcock, Curtis E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, *118*.