



Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework

Xingrui Yu, Xiaomin Wu, Chunbo Luo & Peng Ren

To cite this article: Xingrui Yu, Xiaomin Wu, Chunbo Luo & Peng Ren (2017) Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework, GIScience & Remote Sensing, 54:5, 741-758, DOI: [10.1080/15481603.2017.1323377](https://doi.org/10.1080/15481603.2017.1323377)

To link to this article: <https://doi.org/10.1080/15481603.2017.1323377>



Published online: 05 May 2017.



Submit your article to this journal [↗](#)



Article views: 1310



View related articles [↗](#)



View Crossmark data [↗](#)







Citing articles: 36 View citing articles [↗](#)



ARTICLE

Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework

Xingrui Yu ^a, Xiaomin Wu ^a, Chunbo Luo ^b and Peng Ren ^{a*}

^aCollege of Information and Control Engineering, China University of Petroleum (East China), 66 Changjiang West Road, Qingdao 266580, China; ^bDepartment of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, United Kingdom

(Received 29 October 2016; accepted 22 April 2017)

The recent emergence of deep learning for characterizing complex patterns in remote sensing imagery reveals its high potential to address some classic challenges in this domain, e.g. scene classification. Typical deep learning models require extremely large datasets with rich contents to train a multilayer structure in order to capture the essential features of scenes. Compared with the benchmark datasets used in popular deep learning frameworks, however, the volumes of available remote sensing datasets are particularly limited, which have restricted deep learning methods from achieving full performance gains. In order to address this fundamental problem, this article introduces a methodology to not only enhance the volume and completeness of training data for any remote sensing datasets, but also exploit the enhanced datasets to train a deep convolutional neural network that achieves state-of-the-art scene classification performance. Specifically, we propose to enhance any original dataset by applying three operations – flip, translation, and rotation to generate augmented data – and use the augmented dataset to train and obtain a more descriptive deep model. The proposed methodology is validated in three recently released remote sensing datasets, and confirmed as an effective technique that significantly contributes to potentially revolutionary changes in remote sensing scene classification, empowered by deep learning.

Keywords: deep learning; remote sensing scene classification; convolutional neural network (CNN); big data; data augmentation

1. Introduction

1.1. Background

Scene image analysis has been an important topic in the research literature of remote sensing. Lots of efforts have been made in the low-level analysis such as scene image pixel fusion (Zhou and Gao 2014) and the middle-level analysis such as scene feature extraction and estimation (Tomas et al. 2016). Recently, as the volume of accessible remote sensing image data increases tremendously, the high-level scene image analysis such as scene image classification has attracted especial research interest. There are two types of image classification tasks in the remote sensing literature: (a) pixel-level classification, i.e. to classify pixels in a remote sensing image, and (b) image-level classification, i.e. to classify individual images in a remote sensing image dataset. The

*Corresponding author. Email: pengren@upc.edu.cn

first type of image classification task aims to assign categorical labels to pixels within an image, and the processing can be done in terms of subpixel-based, pixel-based, or pixel-subset-based (e.g. object region) operations. Early pixel-level classification studies include the fuzzy supervised classification (Wang 1990) for pixel categorization. Blaschke *et al.* (2010) presented an overview of object-based image classification methods. Maulik and Chakraborty (2012) proposed a novel semi-supervised supporting vector machine for pixel classification of remote sensing imageries. Myint *et al.* (2011) developed a multi-scale object-based method for pixel-level image classification. Zhang, Chen, and Lu (2015) proposed a subpixel-based approach using linear spectral mixture analysis to detect fractional land cover changes in arid and semiarid urban landscapes. Hussain and Shan (2016) presented an object-based urban land cover classification approach combining aerial digital images and evaluation data. Han and Zhou (2017) developed an adaptive unimodal subclass decomposition learning system for land use classification. Piazza *et al.* (2016) demonstrated the potential of object-based classification for mapping and discrimination of tropical successional forest stages. Chu *et al.* (2016) developed an efficient multi-sensor data fusion approach that integrates full-waveform LIDAR and hyperspectral data to enhance tea and areca classification. Maclaurin and Leyk (2016) reported that information extraction using active learning maximum entropy classification method can be used for effective temporal updating of land cover data. The second type of image classification aims at classifying each scene image into a scenic category, e.g. river, forest, etc., and our work focuses on this type of classification. One widely accepted way to develop such classification algorithms is to train a certain classification model given sufficient training data, i.e. remote sensing images with known scene class labels. The trained model is then used for estimating the scene class labels of unknown remote sensing images. In this regard, various machine learning models, e.g. spectral mixture analysis (Tang and Pannell 2009), multi-feature fusion probabilistic topic model (Zhong, Zhu, and Zhang 2015), and sparse coding (Cui, Schwarz, and Datcu 2015), have been exploited for classifying remote sensing images.

Evidenced in these studies, the quality and quantity of training data have escalated to pave the pathway for more sophisticated high-performance pattern analysis and recognition techniques, among which *deep learning* is particularly promising and has proved to provide an effective means to learn hierarchical representations from large volumes of image data (LeCun, Bengio, and Hinton 2015).

Emerging from the classical machine learning domain, *deep learning* constructs learning models with multiple processing layers that have the ability of hierarchically representing features of raw data. The deep learning methods have triumphed in tackling many pattern recognition and machine learning challenges that were deemed to be difficult (LeCun, Bengio, and Hinton 2015). One key reason for the effectiveness of deep learning is that one complicated deep model can be properly fitted by sufficiently big data such that the diversity and variability of the training data are comprehensively characterized. One of the most popular datasets for training deep models for normal image analysis is ImageNet (Russakovsky *et al.* 2015), which consists of some 15 million labeled images from 22,000 classes. Benefited from such big data, deep learning models have shown great power in normal image analysis tasks such as detection, super-resolution, segmentation, and classification.

Typical deep learning models include deep belief networks (DBNs) (Hinton, Osindero, and Teh 2006) and convolutional neural networks (CNNs) (Hubel and Wiesel 1962). One DBN is a network stacked by restricted Boltzman machines that

are first pre-trained in a layer-wise manner and then finely tuned through back propagation. Zou et al. (2015) introduced DBNs to the remote sensing community by developing a DBN-based support vector machine for scene classification. Basu et al. (2015), from NASA Ames Research Center, proposed a statistical feature extraction method for training a DBN and compared the performance of alternative deep nets for remote sensing scene classification. On the other hand, one CNN stacks a network using interchanged convolutional filtering and pooling, which can be applied to raw images straightforwardly through extracting features hierarchically and classifying the features via the final fully connected layer. Equipped with such advantages, CNNs have been widely used in various image analysis scenarios such as text recognition (Wang et al. 2012) and face identification (Sun, Wang, and Tang 2014). In order to exploit the capability of CNNs for remote sensing scene classification, Zhang, Du, and Zhang (2015) proposed a gradient boosting CNN which outperforms the scene classification schemes based on classical machine learning methodologies.

Different from existing remote sensing image classification methods which focus on improving (deep) machine learning algorithms, we investigate how to increase the diversity of training data. The volume and diversity of training data are essentially important in training a robust deep learning model. It has been observed that one deep model trained based on data with sufficient diversity tends to outperform the same model trained based on data with limited variability (Hinton et al. 2012). This observation reflects the necessity of data augmentation for scene classification, especially in the situation of limited available labeled remote sensing images in contrast to the increasing amount of remote sensing data. One early data augmentation method was studied by Simard, Steinkraus, and Platt (2003), which proposed label-preserving transformations. Recent data augmentation approaches include basic reformation of original images such as cropping and stretching (Dieleman, Willett, and Dambre 2015). Krizhevsky, Sutskever, and Hinton (2012) proposed a data augmentation method to alter intensities of the Red, Green and Blue (RGB) channels of raw data and achieved improved performance on the ImageNet benchmark. In our work, we demonstrate how to exploit data augmentation as a preprocessing procedure for training a deep CNN and empirically evaluate the effectiveness of our data augmentation strategy for lifting the CNN representational power.

1.2. Motivation and contributions

As discussed in Section 1.1, the representational power of one deep model highly relies on the diversity of training data. However, state-of-the-art deep learning strategies in remote sensing mainly focus on designing novel multilayer representations, and are yet to investigate the impact of size and diversity of the training dataset toward their performance. The deficiency of suitable training data in remote sensing is a significant obstacle for realizing the full power of deep learning. For example, one largest labeled remote sensing dataset SAT-4 has only 500,000 images, and its volume sharp contrasts that of the popular normal image dataset ImageNet which contains 15 million labeled images from 22,000 classes.

This article aims to address these fundamental data limitations that hinder the maximization of deep learning's power in remote sensing image classification. We introduce a methodology to enhance the volume and diversity of remote sensing datasets, and exploit the enhanced datasets to train a deep CNN. In order to maintain the key

original feature representations and avoid distortions, we carefully select three basic data augmentation operations including flip, translation, and rotation, which not only significantly enhance the size and completeness of the dataset, but also preserve the scene topologies. We further investigate the application of the enhanced dataset in training one deep CNN and the impact of its performance in remote sensing scene classification. In our implementation, the augmentation approach adopts the set of simple operations with low computational complexity, and is performed on the central processing unit (CPU), while the training is conducted on graphics processing units (GPUs). Our contributions to the remote sensing literature are twofold. Methodologically, we introduce a data augmentation strategy to diversify the training dataset, lifting the representation power of normal CNNs for scene classification. Empirically, experimental results achieve state-of-the-art performance on benchmark remote sensing imagery datasets and outperform existing scene classification deep models.

1.3. Advantages of deep learning models for remote sensing image processing

Classification of images with unique land cover types has attracted increasing research interest in recent years. There are several benchmark datasets dedicated to this purpose of study, and representatives include the UC Merced Land Use, RSSCN7 datasets, etc. Furthermore, NASA has recently released two datasets, SAT-4 and SAT-6. The common premise of all these datasets is that each image is associated with one specific land cover type. These datasets not only enable the possibility of comprehensively training classification algorithms for scene classification, but also provide a common ground for empirically evaluating the algorithms.

Remote sensing scene classification is a challenging task, because scene images may exhibit ambiguous diversities across different types and have sophisticated intra-class variation. For example, certain images from different land cover types have very similar visual appearances. On the other hand, some images belonging to the same type can exhibit considerable visual differences. Such paradoxical phenomena can be observed from the scene image examples in Figure 1, which exhibit contrastive visual variabilities.

The images in Figure 1(a) include two different land cover types: the left two images belong to “Grass”, while the right two images belong to “Field”. However, as the categories “Grass” and “Field” have intrinsic resemblance, the four images appear to associate one common land cover type. On the other hand, the images in Figure 1(b)



Figure 1. Image samples from the RSSCN7 dataset for illustrating the ambiguities in remote sensing scene classification tasks: (a) small inter-class differences – the left two image samples from “Grass” are quite similar to the right two image samples from “Field” and thus tend to be misclassified into one category, and (b) big intra-class variability – the four image samples from “Resident” appear quite different from one another and may be misclassified into different categories.

belong to the same “Resident” category, but they exhibit considerable visual differences and may be misclassified into different categories.

Therefore, to develop learning models that can accurately characterize these complicated visual variations and resemblance is a challenging task that has received significant research interest. A comprehensive review of existing scene classification techniques as well as deep learning methods is given by Xia et al. (forthcoming). Their comparison experimental results reveal that the existing deep learning models such as GoogleLeNet, VGG-VD-16, and CaffeNet, which did not employ the data augmentation strategies, outperform the traditional alternative classification methods. In our article, we will show that the data augmentation strategies are capable of further increasing the representational power of deep CNNs and achieve state-of-the-art performance.

2. The data augmentation enhanced deep learning framework

In this section, we introduce the data augmentation enhanced deep learning methodology for remote sensing scene classification. We first describe the basic data augmentation operations for enhancing the volume and diversity of a dataset, then present a deep CNN exploiting the enhanced dataset, and finally explain how to train the network using the augmented dataset.

2.1. Data augmentation

The representational power of machine learning models (especially deep learning models) highly relies on the training procedures by using plenty of diverse training data. However, though the amount of remote sensing data keeps increasing every year, the properly labeled remote sensing images available for training a deep machine learning model are still limited. We describe how to enhance existing remote sensing datasets via data augmentation and produce augmented datasets to train a more robust deep CNN.

Data augmentation aims at generating additional and more diversified data samples through certain transformations conducted upon original data. In our work, we use publicly available remote sensing scene image sets as the original datasets. For a given remote sensing scene image set $\mathcal{D}_o = \{I_1, \dots, I_K\}$, where I_k indicates the k th image sample in the dataset. Suppose that I_k has totally N pixels. The pixel homogeneous coordinate matrix \mathcal{P}_k for I_k is

$$\mathcal{P}_k = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{bmatrix} \quad (1)$$

where each row represents the homogeneous coordinate for one pixel.

The data augmentation operation on one image I_k is to apply an affine transformation matrix \mathcal{M} to its homogeneous coordinate matrix \mathcal{P}_k and obtain a transformed homogeneous coordinate matrix \mathcal{P}_k^t for the image. The operation is presented as follows:

$$\mathcal{P}_k^t = \mathcal{P}_k \mathcal{M} \quad (2)$$

Here each row of \mathcal{P}_k^t is the transformed homogeneous coordinate for one pixel.

Table 1. Transformation matrices for augmentation operations.

Operations	Flip	Translation	Rotation
Transform matrix	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ T_x & T_y & 1 \end{bmatrix}$	$\begin{bmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix}$

There are various ways to determine the affine transformation matrix \mathcal{M} . In our work, we use three types of random perturbations for generating new augmented data. The three types of transformations are described as follows:

- **Flip** (denoted as \mathcal{T}_1): The image is flipped along the horizontal dimension. The corresponding affine transformation matrix \mathcal{M} is shown in the “Flip” column of Table 1.
- **Translation** (denoted as \mathcal{T}_2): The image is shifted in both the x and y directions of the image. The corresponding affine transformation matrix \mathcal{M} is shown in the “Translation” column of Table 1. T_x and T_y are the offsets on the coordinate axis.
- **Rotation** (denoted as \mathcal{T}_3): The image is rotated with an angle sampled from 0° to 180° . The corresponding affine transformation matrix \mathcal{M} is shown in the “Rotation” column of Table 1, where β is the rotation angle.

Given one image I_k , the augmented data is denoted as $O_k = \{\mathcal{T}_1(I_k), \mathcal{T}_2(I_k), \mathcal{T}_3(I_k)\}$. The augmented dataset for the original dataset \mathcal{D}_o is denoted as $\mathcal{D}_a = \{O_1, \dots, O_K\}$. The augmentation process of the dataset \mathcal{D}_o is thus formulated as follows:

$$\mathcal{D}_a = \{O_1, \dots, O_K\} = \bigcup_{k=1}^K \bigcup_{i=1}^3 \mathcal{T}_i(I_k) \quad (3)$$

The augmented dataset \mathcal{D}_a , along with the corresponding class labels, is then used to train a CNN for the purpose of remote sensing image classification.

It should be noted that we exploit flips, translations, and rotations as the data augmentation operations because they do not change the scene topologies in remote sensing imageries, which is essentially important for consistent scene classification. These operations do not increase the spectral or topological information for the data. On the other hand, the data augmentation (i.e. flips, translations, and rotations) for one individual image diversifies its holistic spatial layout and orientation subject to topological preservation. Considering individual scene images as data samples for classification, the data augmentation enhances the intra-class data diversity and does not incur inter-class ambiguities.

One advantage of deep CNNs over common machine learning methods is their greater capability of characterizing the immense diversity of big data. The augmentation operations such as flips, translations, and rotations increase the diversity of training data, and are thus able to help the deep model capture the data intrinsics more comprehensively. The representational power of a deep CNN can be greatly improved through the comprehensive training based on augmented data than that without data augmentation.

The same data augmentation operations do not necessarily improve the classification accuracy for traditional machine learning methods, which are usually “shallowly”

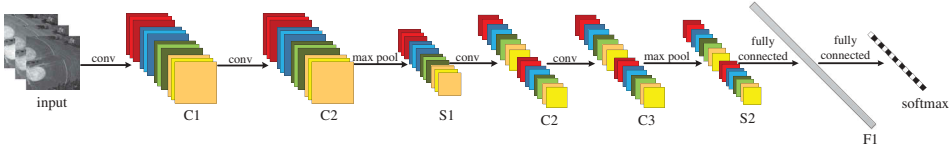


Figure 2. The CNN architecture for remote sensing scene classification. It is composed of four convolutional layers (C1–C4), two pooling layer (S1–S2), one fully connected layer (F1), and one softmax layer (softmax). It is the baseline model of CNN for remote sensing scene classification. For full color versions of the figures in this article, please see the online version.

structured with less parameters and thus have weaker capabilities than deep models in terms of characterizing data diversity.

2.2. Convolutional neural network

CNNs are a type of feed-forward artificial neural networks, with the multilayer structure consisting of convolutional layers, pooling layers, and fully connected layers. We use the CNN structure illustrated in Figure 2 for remote sensing scene classification. Here, a remote sensing image is one input of the CNN. C1, C2, C3, and C4 are four convolutional layers. Each convolutional layer consists of feature maps generated by applying convolutional kernels to convolving the previous layers. S1 and S2 are two max-pooling layers. The max-pool operation maps the maximum within one local region to one number. One max-pooling layer consists of feature maps obtained from applying max-pool operations to previous layer feature maps. Different colors in Figure 2 indicate feature maps generated with different convolutional kernels. The CNN is finalized by a fully connected layer and a softmax layer which outputs the classification result. We refer the interested readers to the landmark work (Simonyan and Zisserman 2015) for CNN construction details.

2.3. Training a CNN

Training a CNN is to compute the optimal parameter values for the network based on a training set, i.e. labeled remote sensing images as inputs and their corresponding labels as target outputs. In our work, we use the augmented dataset \mathcal{D}_a described in Section 2.1 as the training input. One input image is processed through sequential interchanged convolution and max-pooling layers. Each convolution layer generates feature maps and each max-pooling layer downsizes the feature maps in terms of the neighboring maximization pooling strategy. A convolution and max-pooling couple is illustrated in Figure 3.

The layers C1, C2, and S1 (or similarly C3, C4, and S2) in Figure 2 can be thought of examples of the $(n - 1)$ th, n th, and $(n + 1)$ th layers in a concrete CNN, respectively.

Specifically, the feature maps for the n th layer, which is a convolution layer, are obtained by convolving the $(n - 1)$ th layer with trainable parameters (i.e. the weight W_n and bias b_n) and then being processed by an activation function $f(\cdot)$. The trainable parameters are initialized randomly subject to a uniform distribution. The convolution and activation operations (marked in red in Figure 3) result in the feature maps for the n th layer as follows:

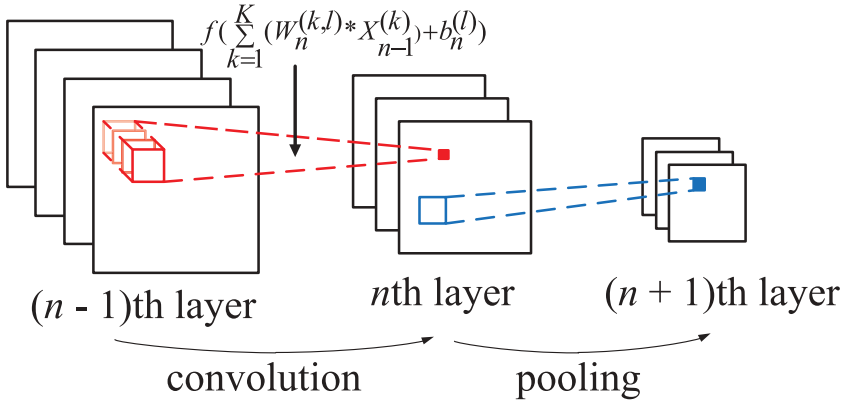


Figure 3. A convolution and max-pooling couple. The convolutional layer (i.e. the n th layer) is obtained by convolving the input $X_{n-1}^{(k)}$ (i.e. the $(n-1)$ th layer) with a set of trainable filters $W_n^{(k,l)}$ and then being processed by an activation function $f(\cdot)$ (e.g. LReLU). The max-pooling layer (i.e. the $(n+1)$ th layer) aims to reduce the spatial size of feature maps (i.e. the n th layer) by outputting the maximal values of spatial local regions on the feature maps and hence to reduce computation.

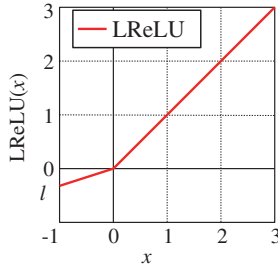


Figure 4. An example of the LReLU nonlinearity function.

$$X_n^{(l)} = f\left(\sum_{k=1}^K (W_n^{(k,l)} * X_{n-1}^{(k)}) + b_n^{(l)}\right) \quad (4)$$

Here, the activation function $f(\cdot)$ is a leaky rectified linear unit (LReLU) (Maas, Hannun, and Ng 2013) as illustrated in Figure 4, and $*$ indicates the convolution operation.

The K feature maps for the $(n-1)$ th layer is represented as a set of matrices $X_{n-1}^{(k)}$ ($k = 1, 2, \dots, K$), and the l th feature map for the n th layer is represented as $X_n^{(l)}$. A max-pooling layer (e.g. the $(n+1)$ th layer in Figure 3) maps maxima within local regions of the previous layer to individual numbers (marked in blue in Figure 3). The feature maps are finally convolved with fully connected layers and generate a target output (i.e. a predicted class label). We measure the error between the target output and the true label of the input remote sensing image. The trainable parameters for each layer are optimized subject to the error minimization. The optimization of trainable parameters is achieved by an effective back propagation method referred to as the mini-batch stochastic gradient descent (Ngiam et al. 2011). We repeatedly optimize the trainable parameter throughout

all augmented training data and obtain the trained network. Specifically, the network utilized in our work is one simplified version of VGGNet (Simonyan and Zisserman 2015) with a structure of eight layers (as illustrated in Figure 2). To prevent overfitting in training the CNN, a dropout operation (Srivastava et al. 2014) is applied after every max-pooling layer. Specifically, the implementation of dropout involves randomly choosing a certain number of neurons (i.e. activation function units) during each training step and performing back propagation only through them. Dropout is a regularization technique for training networks, which prevents CNNs from overfitting (Srivastava et al. 2014).

The trained network is then used for classifying unlabeled remote sensing images by predicting class labels for them. For an unlabeled remote sensing image, we predict its class label by using the trained convolution neural network with the optimal parameters. In testing, the convolution, activation, and pooling are performed in similar ways as those in training. The only difference is that the parameters such as W_n and b_n in Equation (4) are adjustable in training. However, they are fixed optimal values in testing. It is in such a way that the label of an unknown remote sensing image is predicted.

3. Experimental evaluations

In this section, we empirically evaluate our strategy for training a deep CNN for scene classification based on data augmentation. We first introduce the benchmark remote sensing image datasets, then describe the experiment settings, and finally present the experimental results of alternative methods on the benchmark datasets.

3.1. Datasets

Three benchmark remote sensing datasets are used for experimental evaluations. The first dataset is SAT which contains two sub-datasets, i.e. SAT-4 and SAT-6 datasets (<http://csc.lsu.edu/~saikat/deepsat/>). The second dataset is RSSCN7 (<https://sites.google.com/site/qinzoucn/documents>). The third dataset is UC Merced Land Use (<http://vision.ucmerced.edu/datasets/landuse>).

Both SAT-4 and SAT-6 were extracted from the NASA National Agriculture Imagery Program dataset. Specifically, SAT-4 consists of 500,000 images which cover four scene classes, i.e. barren lands, trees, grasslands, and a class involving various scenes other than the above three. SAT-6 consists of 405,000 images which cover six scene classes, i.e. barren lands, trees, grasslands, roads, buildings, and water bodies. The resolution for individual images in SAT-4 and SAT-6 is 28×28 . The RSSCN7 dataset contains 2800 remote sensing images which are from seven typical scene categories, i.e. grasslands, forests, farmlands, parking lots, residential regions, industrial regions, and water bodies. For each category, there are 400 images sampled on four different scales with 100 images per scale. The resolution of individual images is 400×400 . The RSSCN7 dataset is rather challenging due to the wide differences of the scene images which were captured under changing seasons and varying weathers and sampled with different scales. The UC Merced Land Use dataset is a popular dataset for remote sensing scene classification and contains 2100 images which are from 21 scene categories. For each category, there are 100 images. The resolution for individual images is 256×256 . Images in UC Merced Land Use were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country.

Table 2. Candidate parametric values of augmentation parameters and CNN hyper-parameters for grid search.

Datasets	Augmentation parameters			CNN hyper-parameters
	Rotation (degree)	Width shift (degree)	Height shift (degree)	Learning rate (value)
SAT	10, 20, 30	0.1, 0.2, 0.3	0.1, 0.2, 0.3	0.005, 0.001
RSSCN7	10, 20, 30	0.1, 0.2, 0.3	0.1, 0.2, 0.3	0.0001, 0.0005
UC Merced Land Use	10, 20, 30	0.1, 0.2, 0.3	0.1, 0.2, 0.3	0.001, 0.005

3.2. Experiment settings

To validate the effectiveness of the data augmentation, we design two sets of experiments for training one common CNN. The first set of experiments uses the original datasets to train the CNN. The second set of experiments first performs the data augmentation to the original datasets and then uses the augmented datasets to train the same CNN. We refer to the first and second set of experiments as non-aug-experiments and aug-experiments, respectively. For aug-experiments, we configure the augmentation parameters and the CNN training hyper-parameters by searching optimal parametric values from a grid set of candidate parametric values, which include three rotation degree candidates, three width shift ratio candidates, three height shift ratio candidates, and two learning rate candidates. The candidate parametric values for grid search with respect each dataset are described in Table 2. The mathematical relations between the augmentation parameters and the parameters in the transformation matrices in Table 1 are presented in Equation (5).

$$\begin{cases} \beta = \pi * \mathcal{U}(-\text{Rotation}, \text{Rotation}) / 180 \\ T_y = \mathcal{U}(-\text{Widthshift}, \text{Widthshift}) * W \\ T_x = \mathcal{U}(-\text{Heightshift}, \text{Heightshift}) * H \end{cases} \quad (5)$$

where $\mathcal{U}(a, b)$ denotes a random value sampled subject to the uniform distribution in the interval $[a, b]$, and W and H denote the width and height of one image, respectively.

It is clear that we totally have 54 sets of parametric configurations for training a CNN based on an augmented dataset. In order to observe the effects of data augmentation for training a CNN model, we split subsets of SAT-4 and SAT-6 and the whole datasets RSSCN7 and UC Merced Land Use into training, validation, and test subsets separately, as described in Table 3. For each dataset, we conduct 54 individual aug-experiments and one non-aug-experiment.

Table 3. Dataset settings for experimental evaluation.

Datasets	Subsets (number of images)			Image resolution	
	Training	Validation	Test	Height	Width
SAT-4	2800	1199	1000	28	28
SAT-6	2268	971	810	28	28
RSSCN7	1400	700	700	400	400
UC Merced Land Use	1050	630	420	64	64

Especially, each of the 54 individual aug-experiments is set subject to a combination of parametric candidates (i.e. three rotation candidates, three width shift candidates, three height shift candidates, and two learning rate candidates). The experimental results are compared in terms of Kappa index (i.e. Cohen's Kappa) (Hrechak and Mchugh 1990; Cohen 1960).

The data augmentation operations are performed on CPU, and the CNN training procedures are conducted on an NVIDIA GeForce GTX TITAN X 12 GB GPU. For all aug-experiments, we trained the deep models with the same architecture as described in Sections 2.2 and 2.3 based on the augmented datasets, and for all non-aug-experiments, we train the same deep models based on the original datasets.

3.3. Quantitative experimental evaluations

We train a CNN based on each of the four datasets, i.e. SAT-4, SAT-6, RSSCN7, and UC Merced Land Use, separately. For each dataset, both aug-experiments and non-aug-experiments are conducted. The parametric values for the aug-experiments are set according to Table 2. We thus have 54 different sets of parametric values for training a CNN using one augmented dataset. The testing results in terms of Kappa index are shown in Table 4. The “Non-aug” column gives the Kappa indices for testing the CNN trained by using original datasets. On the other hand, each entry of the “Aug” column gives a Kappa index range for testing the CNN trained by using the augmented dataset subject to the 54 different sets of parametric configurations. It is clear that for each dataset, even the smallest Kappa index for aug-experiments is greater than that of non-aug-experiments. We perform grid search to obtain optimal parametric configurations, which are shown in Table 5. These experiments validate that one CNN trained by using augmented remote sensing dataset outperforms that trained by using the original remote sensing dataset.

Table 4. Experimental results for non-aug-experiment and aug-experiments on the four datasets in terms of Kappa index.

Datasets	Kappa index	
	Non-aug	Aug
SAT-4	0.83	0.87–0.96
SAT-6	0.94	0.94–0.97
RSSCN7	0.61	0.71–0.86
UC Merced Land Use	0.48	0.71–0.87

Table 5. Optimal parametric settings for data augmentation and CNNs for aug-experiments subject to Kappa index.

Datasets	Augmentation parameters			CNN hyper-parameters	
	Rotation	Width shift	Height shift	Learning rate	Kappa index
SAT-4	10	0.3	0.1	0.005	0.96
SAT-6	10	0.3	0.3	0.005	0.97
RSSCN7	20	0.1	0.3	0.0001	0.86
UC Merced Land Use	10	0.2	0.3	0.005	0.87

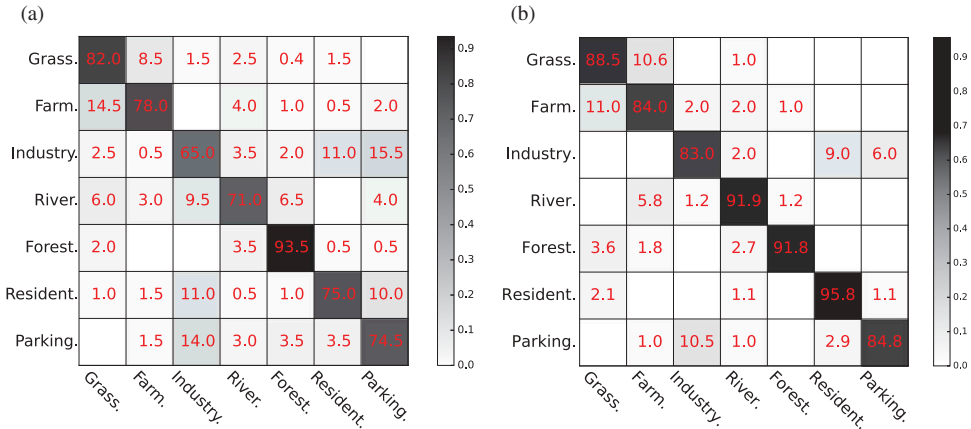


Figure 5. Confusion matrices on the RSSCN7 dataset: (a) the deep feature selection model (Zou et al. 2015) and (b) our data augmentation enhanced CNN model.

To make the empirical evaluation one step further, we experimentally compared our framework with two state-of-the-art deep-learning-based scene classification approaches.

We first test our model with the augmented data and compare it with the state-of-the-art deep feature selection model (Zou et al. 2015) using the RSSCN7 dataset. We use 50% of the data for training our CNN and use the “Test” subset for testing the classification performance of our trained model. Our method achieves a Kappa index of 0.86, which is better than the 0.73 obtained by the deep feature selection model (Zou et al. 2015). Figure 5 shows the comparison of the confusion matrices of the classification results achieved by the deep feature selection model and our method. It is clear that the classification accuracy of our model outperforms the state-of-the-art deep feature selection model on all categories except *forest*. Here, we observe that the classification accuracy of the deep feature selection model has large variation over different categories. On the other hand, the classification accuracy of our model in each category is not substantially different from the rest. This result reveals that our method not only achieves better overall classification performance but also exhibits greater robustness than the deep feature selection model.

Then, we experimentally test our model based on the datasets SAT-4 and SAT-6. In order to evaluate the fitness of our scene classification model, we draw learning curves during the whole training and testing procedures with 50 epochs. Figure 6 illustrates the training accuracy and testing accuracy of the CNN on the SAT-4 and SAT-6 datasets. We observe that the training curve and test curve fit well with each other. This implies that the data augmentation operations enable us to train a deep model with a reasonable balance between the variance in training and the bias in testing.

In addition, we compare our model with a state-of-the-art deep learning framework DeepSat (Basu et al. 2015), which utilizes 22 features selected from 150 extracted features using feature-ranking to train a DBN classifier. To make a fair comparison, we follow the same experimental setting with DeepSat for our method and do not use our own experimental setting described earlier. The results of DeepSat and our method are shown in Table 6. The classification accuracy of our framework on SAT-4 and SAT-6 datasets reaches 99.127% and 99.297%, respectively, both of which are better than the results obtained by the DeepSat framework.

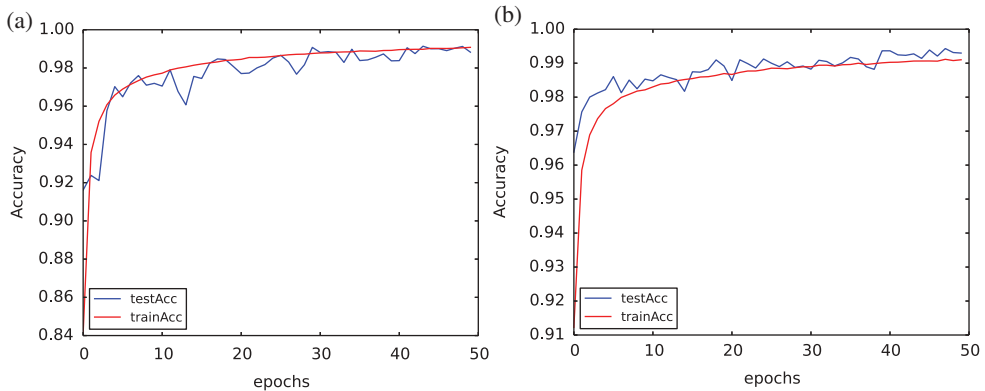


Figure 6. Training and testing accuracy curves of our data augmentation enhanced CNN model with respect to number of epochs. (a) Training and testing curves on SAT-4 and (b) training and testing curves on SAT-6.

Table 6. Performance comparison of DeepSat (Zou et al. 2015) and our data augmentation enhanced CNN model on SAT-4 and SAT-6 datasets in terms of classification accuracy(%).

Method	Classification accuracy (%)	
	SAT-4	SAT-6
DeepSat (Basu et al. 2015)	97.946	93.916
Our method	99.127	99.297

3.4. Qualitative experimental evaluations

In order to qualitatively evaluate the effectiveness of our data augmented CNN model in comparison with the original CNN model, we visualize scene image samples classified by CNN models without and with data augmentation in Figure 7. The image samples misclassified by the original CNNs are marked with surrounding dash squares, and these samples are correctly classified by the data augmentation enhanced CNNs. These visualized classified image samples indicate that CNNs with data augmentation have stronger classification power than those without data classification.

To make the qualitative evaluation one step further, we visualize the image representations extracted from the F1 layers of CNN models without and with data augmentation. We first compute the F1 representational features for all scene images in the test subsets, and then use the t-SNE algorithm (Laurens and Hinton 2008) to embed the F1 features into a two-dimensional (2-D) space. The embedding results are visualized in Figure 8. Here, each point represents an image sample in the F1 feature space and each color indicates a scene category. We show these 2-D embedding points in colors with respect to their true scene categories. The visualization results are shown in Figure 8. Notably, the image features with data augmentation for different categories (the right column of Figure 8) are more clearly separable than those without data augmentation (the left column of Figure 8). This observation illustrates that one CNN model trained with data augmentation is capable of learning more inter-class discriminative representations than that trained without data augmentation for remote sensing scene classification. Furthermore, the feature distribution

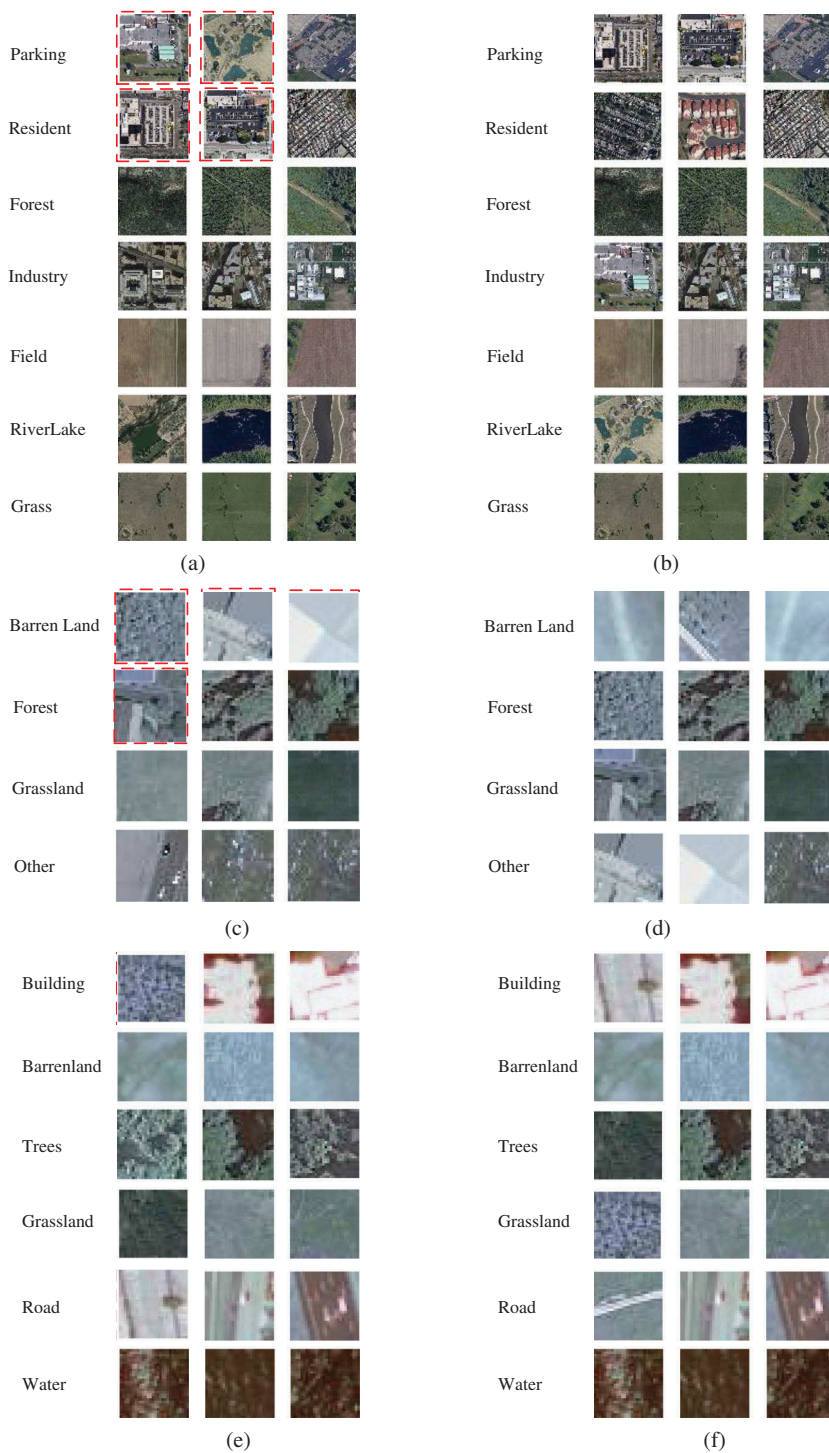


Figure 7. Scene image samples classified by CNN models without and with data augmentation. The visualized scene images are samples from the RSSCN7, SAT-4, and SAT-6 datasets, separately. Misclassified image samples are marked with surrounding dash squares. (a) RSSCN7 without augmentation; (b) RSSCN7 with augmentation; (c) SAT-4 without augmentation; (d) SAT-4 with augmentation; (e) SAT-6 without augmentation; (f) SAT-6 with augmentation.

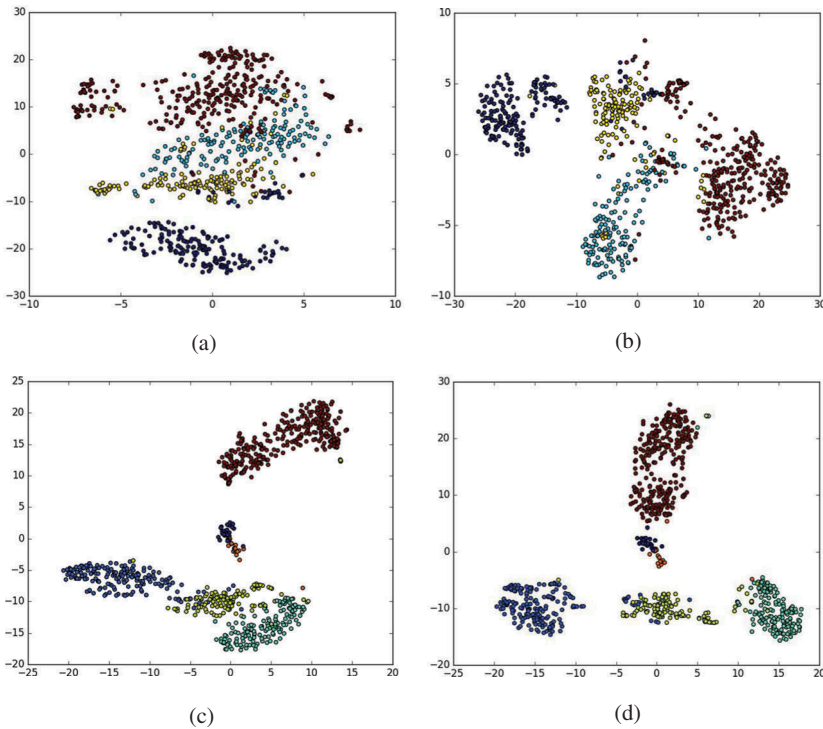


Figure 8. Non-aug and aug F1 feature distributions in terms of t-SNE (Laurens and Hinton 2008) on the SAT-4 and SAT-6 datasets, separately. (a) F1 feature of SAT-4 without augmentation; (b) F1 features of SAT-4 with augmentation; (c) F1 features of SAT-6 without augmentation; (d) F1 feature of SAT-6 with augmentation.

for each scene category in Figure 8(b) and 8(d) exhibits in a consistent form with rare bias. This observation validates the robustness of the data augmentation enhanced CNN in terms of the immunity to intra-class outliers and biases.

4. Conclusions

This article introduces basic data augmentation operations to address the fundamental data limitation problem in applying deep learning for remote sensing image processing. We describe how to use data augmentation to improve the remote sensing scene classification performance of CNNs. We show that the diversity and completeness of data can be greatly enhanced by data augmentation, and when applied to training deep learning models, the experimental results with augmentation operations outperform those from the same deep model architecture training on the original dataset. The effectiveness and robustness of our proposed methodology are confirmed by experiments using practical remote sensing datasets. The proposed methodology advances the state-of-the-art and can significantly contribute to the new horizon of deep learning in remote sensing.

Though the data augmentation strategy enhances the diversity of the dataset to a certain extent, it just increases the visual variability of each training remote sensing

image subject to its intrinsic spectral and topological constraints and does not generate new information for the remote sensing image. Future research will focus on exploiting state-of-the-art generative adversarial nets to generate new remote sensing image instances based on a trained augmentation strategy beyond the basic flip, translation, and rotation operations. Furthermore, based on the effective representational powers of deep models, it is highly possible for the study of remote sensing scene analysis to be extended from categorizing static remote sensing images to analyzing dynamic remote sensing videos.

Acknowledgments

This work was supported by National Natural Science Foundation of China under grant number 61671481; Qingdao Applied Fundamental Research Project under grant number 16-5-1-11-jch; and the Fundamental Research Funds for Central Universities.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Natural Science Foundation of China under grant number 61671481; Qingdao Applied Fundamental Research Project under grant number 16-5-1-11-jch; and the Fundamental Research Funds for Central Universities.

ORCID

Xingrui Yu  <http://orcid.org/0000-0002-8941-2698>

Xiaomin Wu  <http://orcid.org/0000-0002-0898-4185>

Chunbo Luo  <http://orcid.org/0000-0002-9860-2901>

Peng Ren  <http://orcid.org/0000-0003-3949-985X>

References

- Basu, S., S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. R. Nemani. 2015. "DeepSat-A Learning Framework for Satellite Imagery." *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 37: 1–10. doi: [10.1145/2820783.2820816](https://doi.org/10.1145/2820783.2820816).
- Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. doi:[10.1016/j.isprsjprs.2009.06.004](https://doi.org/10.1016/j.isprsjprs.2009.06.004).
- Chu, H. J., C. K. Wang, S. J. Kong, and K. C. Chen. 2016. "Integration of Full-Waveform Lidar and Hyperspectral Data to Enhance Tea and Areca Classification." *Giscience & Remote Sensing* 53 (4): 542–559. doi:[10.1080/15481603.2016.1177249](https://doi.org/10.1080/15481603.2016.1177249).
- Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational & Psychological Measurement* 20 (1): 37–46. doi:[10.1177/001316446002000104](https://doi.org/10.1177/001316446002000104).
- Cui, S., G. Schwarz, and M. Datcu. 2015. "Remote Sensing Image Classification: No Features, No Clustering." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (11): 5158–5170. doi:[10.1109/JSTARS.2015.2495267](https://doi.org/10.1109/JSTARS.2015.2495267).
- Dieleman, S., K. W. Willett, and J. Dambre. 2015. "Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction." *Monthly Notices of the Royal Astronomical Society* 450 (2): 1441–1459. doi:[10.1093/mnras/stv632](https://doi.org/10.1093/mnras/stv632).

- Han, M., and Y. Zhou. 2017. "An Adaptive Unimodal Subclass Decomposition (AUSD) Learning System for Land Use and Land Cover Classification Using High-Resolution Remote Sensing." *Giscience & Remote Sensing* 54 (1): 20–37. doi:10.1080/15481603.2016.1246057.
- Hinton, G., L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, 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. doi:10.1109/MSP.2012.2205597.
- Hinton, G., S. Osindero, and Y. W. Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7): 1527–1554. doi:10.1162/neco.2006.18.7.1527.
- Hrechak, A. K., and J. A. Mchugh. 1990. "Automated Fingerprint Recognition Using Structural Matching." *Pattern Recognition* 23 (8): 893–904. doi:10.1016/0031-3203(90)90134-7.
- Hubel, D. H., and T. N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160 (1): 106–110. doi:10.1113/jphysiol.1962.sp006837.
- Hussain, E., and J. Shan. 2016. "Object-Based Urban Land Cover Classification Using Rule Inheritance over Very High-Resolution Multisensor and Multitemporal Data." *Giscience & Remote Sensing* 53 (2): 164–182. doi:10.1080/15481603.2015.1122923.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25: 2.
- Laurens, V. D. M., and G. Hinton. 2008. "Visualizing Data Using T-Sne." *Journal of Machine Learning Research* 9 (2605): 2579–2605.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444. doi:10.1038/nature14539.
- Maas, A. L., A. Y. Hannun, and A. Y. Ng. 2013. "Rectifier Nonlinearities Improve Neural Network Acoustic Models." *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.
- Maclaurin, G. J., and S. Leyk. 2016. "Temporal Replication of the National Land Cover Database Using Active Machine Learning." *Giscience & Remote Sensing* 53 (2): 759–777. doi:10.1080/15481603.2016.1235009.
- Maulik, U., and D. Chakraborty. 2012. "A Novel Semisupervised SVM for Pixel Classification of Remote Sensing Imagery." *International Journal of Machine Learning and Cybernetics* 3 (3): 247–258. doi:10.1007/s13042-011-0059-3.
- Myint, S. W., P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng. 2011. "Per-Pixel vs. Object-Based Classification of Urban Land Cover Extraction Using High Spatial Resolution Imagery." *Remote Sensing of Environment* 115 (5): 1145–1161. doi:10.1016/j.rse.2010.12.017.
- Ngiam, J., J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng. 2011. "On Optimization Methods for Deep Learning." *International Conference on Machine Learning* 67–105.
- Piazza, G. A., A. C. Vibrams, V. Liesenberg, and J. C. Refosco. 2016. "Object-Oriented and Pixel-Based Classification Approaches to Classify Tropical Successional Stages Using Airborne High Spatial Resolution Images." *Giscience & Remote Sensing* 53 (2): 206–226. doi:10.1080/15481603.2015.1130589.
- Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, and Z. Huang, et al. 2015. "Imagenet Large Scale Visual Recognition Challenge." *International Journal Of Computer Vision* 115 (3): 211–252. doi: 10.1007/s11263-015-0816-y.
- Simard, P. Y., D. Steinkraus, and J. C. Platt. 2003. "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis." *International Conference on Document Analysis and Recognition* 958–962. doi: 10.1109/ICDAR.2003.1227801.
- Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations*.
- Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–1958.
- Sun, Y., X. Wang, and X. Tang. 2014. "Deep Learning Face Representation from Predicting 10,000 Classes." *IEEE Conference on Computer Vision and Pattern Recognition* 1891–1898. doi: 10.1109/CVPR.2014.244.
- Tang, Y., and C. W. Pannell. 2009. "A Hybrid Approach for Land Use/Land Cover Classification." *Giscience & Remote Sensing* 46 (4): 365–387. doi:10.2747/1548-1603.46.4.365.

- Tomas, L., L. Fonseca, C. Almeida, F. Leonardi, and M. Pereira. 2016. "Urban Population Estimation Based on Residential Buildings Volume Using IKONOS-2 Images and Lidar Data." *International Journal of Remote Sensing* 37 (sup1): 1–28. doi:[10.1080/01431161.2015.1121301](https://doi.org/10.1080/01431161.2015.1121301).
- Wang, F. 1990. "Fuzzy Supervised Classification of Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 28 (2): 194–201. doi:[10.1109/36.46698](https://doi.org/10.1109/36.46698).
- Wang, T., J. D. Wu, A. Coates, and A. Y. Ng. 2012. "End-To-End Text Recognition with Convolutional Neural Networks." *International Conference on Pattern Recognition* 3304–3308.
- Xia, G., J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang. forthcoming. "AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification." *IEEE Transactions on Geoscience Remote Sensing*.
- Zhang, C., Y. Chen, and D. Lu. 2015. "Detecting Fractional Land-Cover Change in Arid and Semiarid Urban Landscapes with Ultitemporal Landsat Thematic Mapper Imagery." *Geoscience & Remote Sensing* 52 (6): 700–722. doi:[10.1080/15481603.2015.1071965](https://doi.org/10.1080/15481603.2015.1071965).
- Zhang, F., B. Du, and L. Zhang. 2015. "Scene Classification via a Gradient Boosting Random Convolutional Network Framework." *IEEE Transactions on Geoscience and Remote Sensing* 30 (99): 1–10. doi:[10.1109/TGRS.2015.2488681](https://doi.org/10.1109/TGRS.2015.2488681).
- Zhong, Y., Q. Zhu, and L. Zhang. 2015. "Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 53 (11): 6207–6222. doi:[10.1109/TGRS.2015.2435801](https://doi.org/10.1109/TGRS.2015.2435801).
- Zhou, H., and H. Gao. 2014. "Fusion Method for Remote Sensing Image Based on Fuzzy Integral." *Journal of Electrical and Computer Engineering* 2014: 26–34. doi:[10.1155/2014/437939](https://doi.org/10.1155/2014/437939).
- Zou, Q., L. Ni, T. Zhang, and Q. Wang. 2015. "Deep Learning Based Feature Selection for Remote Sensing Scene Classification." *IEEE Geoscience and Remote Sensing Letters* 12 (11): 2321–2325. doi:[10.1109/LGRS.2015.2475299](https://doi.org/10.1109/LGRS.2015.2475299).