Multimedia Tools and Applications (2019) 78:9667–9689
https://doi.org/10.1007/s11042-018-6548-6

Analysis of the inter-dataset representation ability of deep features for high spatial resolution remote sensing image scene classification

Lijun Zhao [1] · Wei Zhang [1,2] · Ping Tang [1]

Received: 1 January 2018 / Revised: 23 July 2018 / Accepted: 15 August 2018 /
Published online: 27 August 2018

Abstract
Recently, scene based classification has become a new trend for very high spatial resolution remote sensing image interpretation. With the advent of deep learning, the pretrained convolutional neural networks (CNNs) have been proved effective as feature extractors for scene classification tasks in the remote sensing domain, but the potential characteristics and capabilities of such deep features have not been sufficiently analyzed and fully understood. Facing with complex remote sensing scenes with huge intra-class variations, it is still not clear about the limitation of these powerful deep features in exploring essential invariant attributes of remote sensing scenes of the same kind but, in most cases, from separate sources. Therefore, this paper makes an intensive investigation in the feature representation ability of such deep features from the aspect of inter-dataset scene classification of remote sensing images. Four well-known pretrained CNN models and three different commonly used datasets are selected and summarized. Firstly, deep features extracted from various intermediate layers of these models are compared. Then, the inter-dataset feature representation ability is evaluated using cross-classification of different datasets and discussed in terms of imaging spatial resolution, image size, model structure, and time efficiency. Finally, several instructive findings are revealed and conclusions are drawn regarding the strength and weakness of the CNN features in the application of remote sensing image scene classification.

Keywords Remote sensing image · Scene classification · Inter-dataset feature representation · Deep learning features · Convolutional neural networks (CNNs)

✉ Lijun Zhao
zhaolj01@radi.ac.cn

[1] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, People's Republic of China

[2] Univiersity of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

1 Introduction

With the constant development of imaging technology of satellite sensors, the spatial resolution has been greatly improved. These very high spatial resolution (VHSR) remote sensing images can provide more detailed information about the Earth, compared with the traditional middle and low resolution remote sensing images. These ground details usually contain multiple land-cover types and complex ground objects, e.g., airports, parking lots, and residential areas, making the previous homogenous regions in middle or low resolution images become highly heterogeneous. Under such circumstances, traditional pixel-based or object-based land-cover classification methods [2, 14, 43] can no longer meet the demand of the present research of image interpretation. These years, scene based remote sensing image classification has attracted more and more attention from the remote sensing community [6, 7, 9–11, 26, 31, 36, 39, 44, 46–48], which provides a new way to exploit the information of VHSR remote sensing images.

Observe that, in order to do scene classification, feature extraction and representation becomes a key procedure. Inspired by recognition technologies from the computer vision community, earlier studies mainly focus on the well-known bag-of-visual-words (BOVW) model [6, 7, 12, 31, 34, 36, 44, 47–49]. Although these studies have produced good results in VHSR remote sensing image scene classification, no more breakthroughs in classification performance have further been achieved using the BOVW-type methods in recent years, owing to the limitation of description capability of the BOVW model itself.

Recently, with the advent of the deep learning methods [1, 15, 23] which have achieved great success in many practical applications like video retrieval and popularity prediction [8, 27], clothing matching [38], aerial image object detection [32] and hyperspectral image classification [41], the feature representation for image scenes has stepped into a new era. Unlike the traditional hand-crafted features (e.g., the BOVW model), the deep learning model can learn a set of rich nonlinear representations directly from the input data with no assumptions or prior knowledge [25]. Compared with the shallow-structured spatial pyramid matching (SPM) model [20], a popular BOVW-type feature which represents an image in different scales, the deep learning features contain much more powerful feature representations and representative structural information of data with multiple levels of abstraction. Such an advantage makes the deep learning features more competent for exploring the intrinsic attributes of complex remote sensing scenes with huge intra-class variations. In the most recent year, preliminary attempts [17, 20, 24, 25, 30, 42, 45] have been made to apply the deep learning models to deal with the VHSR remote sensing image scene classification tasks. Among these existing studies, the deep convolutional neural networks (CNNs) [22] become the most popular deep learning approach in the image classification field. Owing to their high-powered feature learning and representation abilities, dramatic improvements have been achieved beyond the state-of-the-art records on several benchmark datasets [17, 30]. Penatti, Nogueira, and dos Santos [30] first proposed to evaluate the generalization power of the deep CNNs trained for recognizing everyday objects in the aerial and remote sensing domain. Also, Hu et al. [17] thoroughly investigated the CNN features extracted from different layers and proved that not only the fully-connected layers but also the convolutional layers play an important role in the representation of image scene features.

Although the pretrained CNN models from the natural image domain, regarded as feature extractors, have been proved to be successfully applied to the remote sensing community, the application scenarios usually appear in intra-dataset scene classification, i.e., training and

prediction data are from one same dataset or the same regional satellite images, even though they are separate and non-overlapping. Up till now, few studies have ever been concerned about the inter-dataset representation ability of deep features. Here, the deep feature refers to the feature extracted by some pretrained CNN model and the inter-dataset feature representation ability means the ability of deep features in extracting invariant information of data from a same category but from different datasets. Figure 1 visually shows an example of intra-dataset and inter-dataset feature representation and classification. This study is of great importance and makes great senses because in practical applications, we usually have to recognize new unknown remote sensing scene images based on previously learned scene patterns, and these unknown images may not be always from the same dataset or the same regional satellite image as the training samples, but from a separate data source with different imaging conditions most of the time. As remote sensing scene images are easy to be affected by the changed imaging conditions, e.g., imaging orientations and imaging spatial resolutions, the scene images of the same kind usually show various changes under different imaging conditions. Thus, it is demanded that the deep features be powerful enough to extract the most essential features of scenes from the same category. It naturally becomes a question whether such deep features are insusceptible to the changes of imaging conditions and can capture the intrinsic invariant attributes of scene images of the same kind but from different datasets. However, this question has not been answered by any other references, and it is needed that the performance of such deep features should be further analyzed and investigated in the application scenario of inter-dataset classification where training and prediction data are from different datasets.

Based on this motivation, this paper conducts an intensive analysis of the inter-dataset representation ability of deep features for VHSR remote sensing image scene classification. To this end, four well-known pretrained CNN models are selected as feature extractors, including AlexNet [19], Caffe [18], GoogLeNet [40], and VGG-16 [37]. The selection criteria are that these four models are commonly used in the recent research of remote sensing image scene classification [4, 17, 30] and are also the top two winners in the ImageNet Large Scale Visual

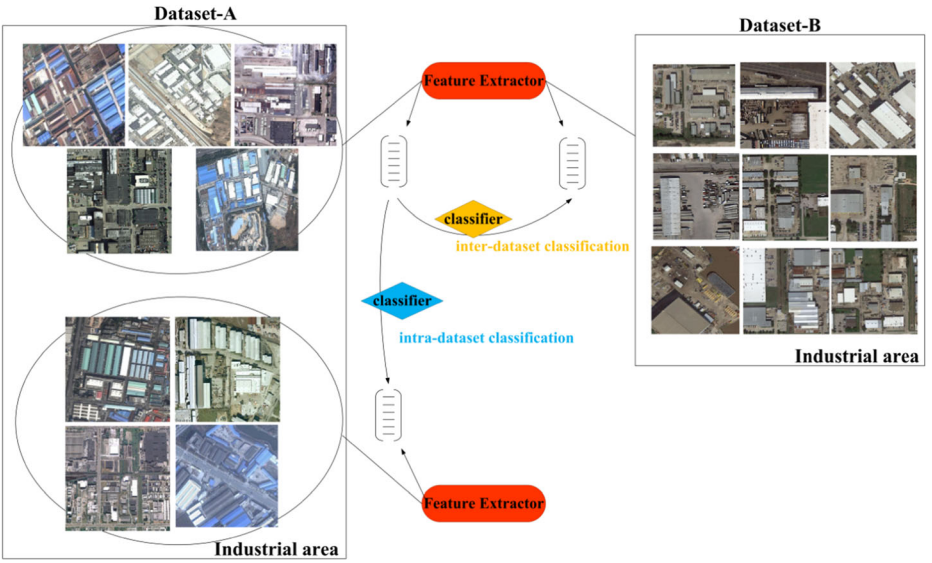

Fig. 1 Illustration of inter-dataset classification and intra-dataset classification

Recognition Challenge (ILSVRC) in recent years. To analyze their inter-dataset representation ability, deep features are evaluated by inter-dataset classification, i.e., the patterns learned from the deep features in one dataset are utilized to distinguish scene images of the same scene categories but from another dataset with variations in image size, imaging angle, orientation, and spatial resolution. Three public VHSR remote sensing scene datasets [13, 44, 49] are used for inter-dataset classification.

The main contributions of this paper are: 1) our work is the first to intensively study and analyze in details the potential characteristics and capabilities of deep CNN features in the application scenario of inter-dataset scene classification of VHSR remote sensing images, which is what previous studies seldom consider and investigate in. Figure 2 shows the difference of previous works and the study in this paper. 2) our results answer the question whether such deep features are insusceptible to the changes of imaging conditions and can capture the intrinsic invariant attributes of scene images of the same kind but from different datasets and reveal their potential in the application of inter-dataset scene classification. The findings provide important significance for high-level feature representation in remote sensing image scene classification and also instructions for later works in further understanding the applicability of CNN feature extractors in remote sensing image scene classification.

2 Materials and methods

2.1 Data

To evaluate the performance of the deep CNN features in VHSR remote sensing image scene classification, three VHSR remote sensing scene image datasets with varied sizes, imaging angles, orientations and spatial resolutions were used to perform the following experiments.

The first dataset (Dataset-1) [44] is a well-known public dataset for remote sensing image scene classification, which is available at http://vision.ucmerced.edu/datasets. This dataset was extracted from United States Geological Survey (USGS) National Map and contains 21 land-use scene categories (Fig. 3), including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection,

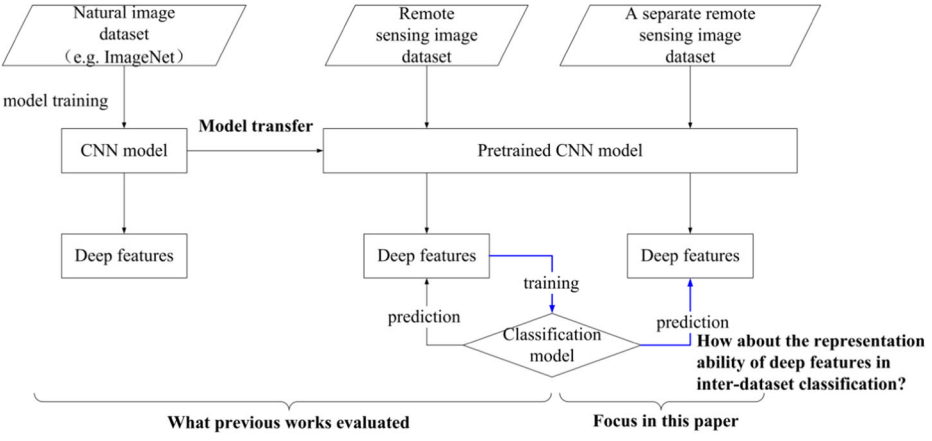

Fig. 2 Comparison of previous works and the study in this paper

medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class includes 100 images and each image has an image size of 256 × 256 pixels. All the images consist of three bands of red, green, and blue with a spatial resolution of about 0.3 m.

The second dataset (Dataset-2) [13] is also a publicly available one, which can be downloaded at http://dsp.whu.edu.cn/cn/staff/yw/HRSscene.html. All the scenes in the dataset were extracted from a set of satellite images exported from Google Earth with spatial resolution up to 0.5 m and three bands of red, green, and blue. The whole dataset contains 19 classes of scenes including airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking lot, pond, port, railway station, residential area, river, and viaduct. For each scene category, there are about 50 scene images, with 1005 images in total for the entire dataset. The image sizes are 600 × 600 pixels. This dataset is a challenging one because all these scenes are extracted from very large satellite images on Google Earth, where the illumination, appearances of objects and

Fig. 3 Examples of Dataset-1

their locations vary significantly, with frequent occlusions [13]. Figure 4 shows some examples of each class in this dataset.

The third dataset (Dataset-3) [49] was created by the authors and has been made available for other researchers, which can be downloaded at http://pan.baidu.com/s/1mhagndY. This dataset was manually extracted from Google Earth, which covers the images of several USA cities including Washington DC, Los Angeles, San Francisco, New York, San Diego, Chicago, and Houston. Three spectral bands were used including red, green, and blue. The spatial resolution is about 0.2 m. All the regional images were cropped into a uniformed image size of 512×512 pixels by an overlapped sampling strategy, and then manually picked out and labeled by the specialists in the field of remote sensing image interpretation for about a week, with those cropped images containing multiple scenes abandoned. Finally, there are 11 complicated scene categories including dense forest, grassland, harbor, high buildings, low buildings, overpass, railway, residential area, roads, sparse forest, and storage tanks. Many of the scene categories are quite similar in vision, which increases the difficulty in distinguishing

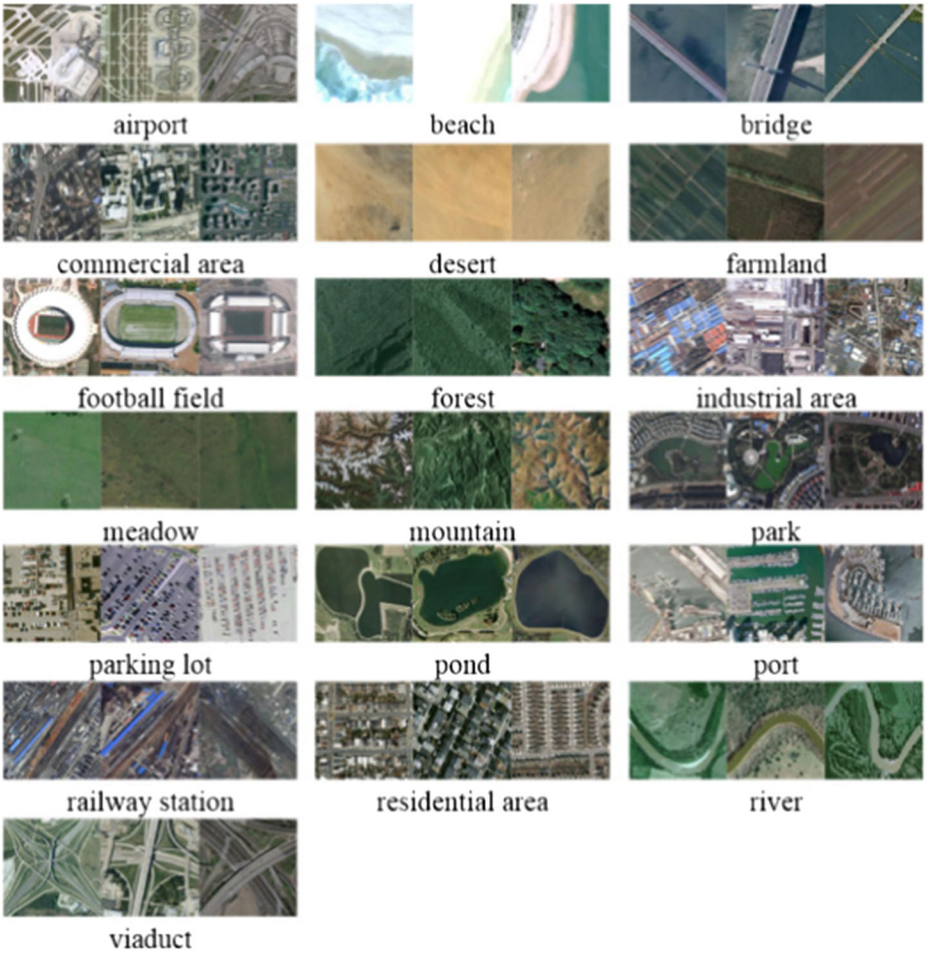

Fig. 4 Examples of Dataset-2

all the scenes. The dataset includes 1232 images in total, with each class about 100 images. Figure 5 shows examples from the dataset.

2.2 Involved deep convolutional neural networks

The CNN model is a type of feed-forward artificial neural network models for deep learning, which is biologically inspired by the organization of the animal visual cortex. Generally, CNN can be considered to be made up of two main parts. The first part contains alternating convolutional and max-pooling layers, in which the convolutional layer outputs feature maps by computing a dot product between the local region in the input feature maps and a filter and the max-pooling layer performs a down-sampling operation to feature maps by computing the maximum on a subregion. As the input of each layer is just the output of its previous layer, a hierarchical feature extractor can be formed to map the original input images into feature vectors. Following the several stacked convolutional and max-pooling layers, the second part contains fully-connected layers, with the last layer (e.g., softmax layer) used to classify the extracted feature vectors. Figure 6 shows an illustration of a typical CNN architecture.

In this paper, pretrained CNN models are used to extract deep CNN features for remote sensing image scene classification. Several architectures have achieved great success in the image classification domain. Here, four successful modern CNN architectures will be briefly depicted and utilized in the following experiments, considering their popularity in image classification tasks. All the pretrained CNN models are utilized from the Caffe Library that is available at https://github.com/BVLC/caffe/wiki/Model-Zoo.

2.2.1 AlexNet

AlexNet, the winning model in the ILSVRC 2012, is developed by Krizhevsky, Sutskever, and Hinton [19]. As shown in Fig. 7, it contains eight layers, in which the first five are

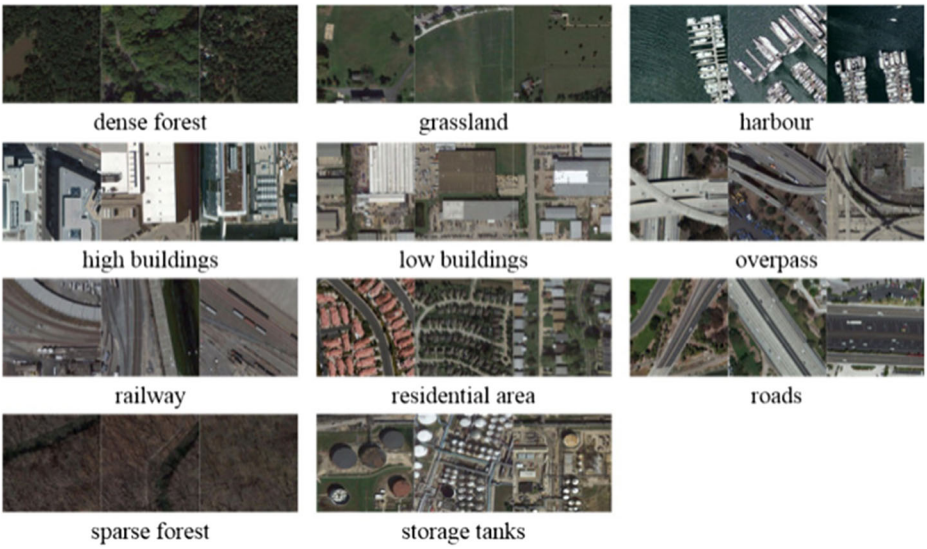

Fig. 5 Examples of Dataset-3

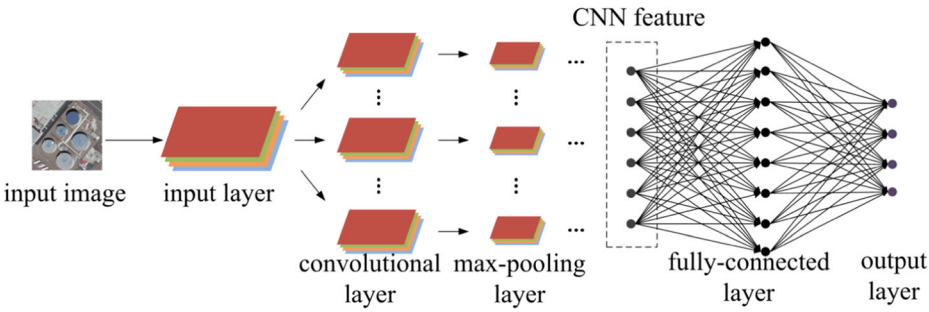

Fig. 6 Illustration of a typical architecture of a CNN model

convolutional and the remaining three are fully-connected, with the output of the last fully-connected layer fed to a 1000-way softmax layer.

Several new characteristics make AlexNet successful in the application of visual recognition tasks, such as rectified linear units (ReLU) nonlinearity, data augmentation, and dropout. To be specific, the ReLU nonlinearity can significantly accelerate the training phase, the data augmentation can effectively combat overfitting by generating image translations and horizontal reflections and altering the intensities of the RGB channels in training images, and the dropout technique can reduce substantial overfitting when used in the first two fully-connected layers.

2.2.2 Caffe

Caffe, also trained in ILSVRC 2012, is maintained and developed by the Berkeley Vision and Learning Centre (BLVC) [18]. It is short for convolutional architecture for fast feature embedding, which can provide multimedia scientists and practitioners with clean and modifiable framework for state-of-the-art deep learning algorithms and a collection of reference models. Similar to AlexNet, it comprises of five convolutional layers and three fully-connected layers, but the order of pooling and normalization layers are exchanged, with no data augmentation in training stage.

2.2.3 GoogLeNet

GoogLeNet, the winning model in ILSVRC 2014, is presented in [40]. The name of GoogLeNet is homage to LeCun's pioneering LeNet-5 network [21]. It is a 22-layer deep

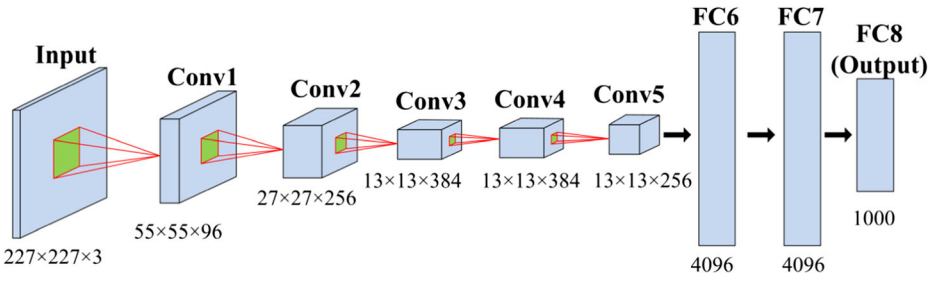

Fig. 7 The overall architecture of AlexNet

CNN when counting only layers with parameters (see Table 1). In contrast to AlexNet, GoogLeNet not only uses much fewer parameters, but also is significantly more accurate.

The core of GoogLeNet is a convolutional neural network architecture, namely the Inception module, which can effectively reduce the complexity of the expensive filters of convolutional architectures by applying dimension reduction and projection, leading to an improved utilization of the computing resources inside the network. As shown in Fig. 8, 1 × 1 convolutions are used to compute reductions before the expensive 3 × 3 and 5 × 5 convolutions. The convolutions with different sizes in the Inception module can produce features at different scales and they are then aggregated and fed to the next layer.

2.2.4 VGG-16

The VGG team secured the first and the second places in ILSVRC 2014 in the localization and classification tasks respectively. Two best-performing deep models, presented by Simonyan and Zisserman [37], were further improved and released after the competition, named as VGG-16 (containing 13 convolutional layers and 3 fully-connected layers) and VGG-19 (containing 16 convolutional layers and 3 fully-connected layers).

Rather than using relatively large receptive fields in the convolutional layers, such as 11 × 11 with stride 4 in the first convolutional layer as in AlexNet, the VGG network uses very small 3 × 3 receptive fields through the whole net. Followed by a stack of convolutional layers are three fully-connected layers, two 4096-channel fully-connected layers and one 1000-channel layer for softmax output. In this work, the pretrained VGG-16 model was selected for feature extraction, since it has fewer layers with competitively high performance. Table 2 shows the architecture of VGG-16 network.

3 Experimental setup

To compare and analyze the inter-dataset representation ability of deep features, the inter-dataset classification was conducted with the support vector machine (SVM) as the unified classifier like many studies [3, 17, 28, 30, 33] that directly use the pretrained CNN models. The flowchart of the feature evaluation procedure is presented in Fig. 9, which consists of two main steps, including feature representation and feature classification.

Table 1 The architecture of GoogLeNet network

Type	Dimension	Depth	Type	Dimension	Depth
Input	224 × 224 × 3	0	Inception(4b)	14 × 14 × 512	2
Convolution	112 × 112 × 64	1	Inception(4c)	14 × 14 × 512	2
Max pool	56 × 56 × 64	0	Inception(4d)	14 × 14 × 528	2
Convolution	56 × 56 × 192	1	Inception (4e)	14 × 14 × 832	2
Convolution	56 × 56 × 192	1	Max pool	7 × 7 × 832	0
Max pool	28 × 28 × 192	0	Inception(5a)	7 × 7 × 832	2
Inception(3a)	28 × 28 × 256	2	Inception(5b)	7 × 7 × 1024	2
Inception(3b)	28 × 28 × 480	2	Average pool	1 × 1 × 1024	0
Max pool	14 × 14 × 480	0	FC	1 × 1 × 1024	1
Inception(4a)	14 × 14 × 512	2	Output	1 × 1 × 1000	0

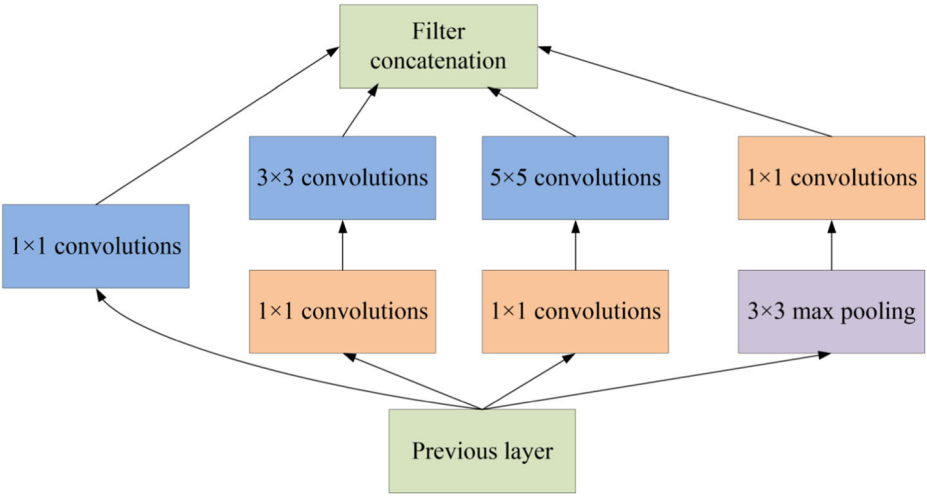

Fig. 8 Inception module with dimension reduction in GoogLeNet

For the deep feature representation, all the involved pretrained CNN models were used as feature extractors and the CNN features were extracted from the intermediate outputs of the convolutional and fully-connected layers. Thus, for the deep features extracted from different layers, the dimension of the obtained feature vector is identical with that of the layer selected. Input images were all resized to the size of ImageNet (i.e., 256 × 256) so as to meet the requirements of the pretrained CNN models. Besides, for comparison purpose, the well-known SPM [20] model was added as the baseline hand-crafted feature extraction method, as it generates features in different scales by partitioning the image into increasingly fine subregions, which is, to some degree, similar to deep learning. Dense regions using a regular grid with 8 pixels spacing and 16 × 16 pixels patch size were used to generate image patches, based on each of which its scale-invariant feature transform (SIFT) feature is represented by histograms of eight-bin gradient directions computed over a 4 × 4 spatial grid. A three-level pyramid was applied as Lazebnik, Schmid, and Ponce [20] to generate increasingly fine subregions. The vocabulary sizes were set to be 100, 300, 500, 700, and 900, and the settings with the best classification results were applied.

For feature classification, the popular LIBSVM toolbox [5] was applied as the unified classification method throughout the experiments. For deep CNN features, the radial basis function (RBF) and linear kernels were applied considering their popularity [17, 29, 30]. For

Table 2 The architecture of VGG-16 network

Type	Dimension	Depth	Type	Dimension	Depth
Input	224 × 224 × 3	0	Conv4_2	28 × 28 × 512	1
Conv1_1	224 × 224 × 64	1	Conv4_3(after pool)	14 × 14 × 512	1
Conv1_2(after pool)	112 × 112 × 128	1	Conv5_1	14 × 14 × 512	1
Conv2_1	112 × 112 × 128	1	Conv5_2	14 × 14 × 512	1
Conv2_2(after pool)	56 × 56 × 128	1	Conv5_3(after pool)	7 × 7 × 512	1
Conv3_1	56 × 56 × 256	1	FC6	1 × 1 × 4096	1
Conv3_2	56 × 56 × 256	1	FC7	1 × 1 × 4096	1
Conv3_3 (after pool)	28 × 28 × 256	1	FC8	1 × 1 × 1000	1
Conv4_1	28 × 28 × 512	1	Output	1 × 1 × 1000	0

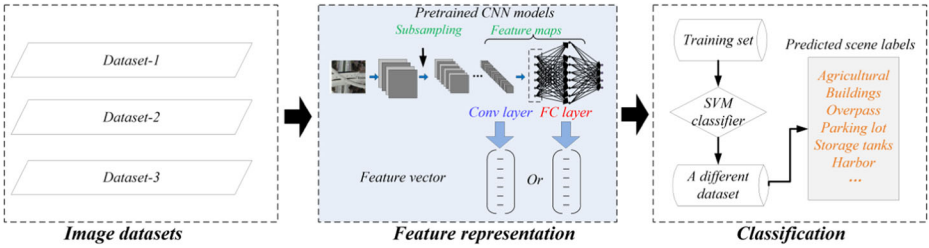

Fig. 9 Flowchart of the comparative study scheme

SPM, the corresponding pyramid match kernel was used as described by Lazebnik, Schmid, and Ponce [20]. The cross-validation method with the grid search mechanism [16] was used for SVM model selection. For the RBF based SVM, the optimal model parameters were obtained to minimize a five-fold cross-validation estimate of the classification error rate using different penalty parameters $C = [2^3, 2^1, 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}, 2^{-9}, 2^{-11}, 2^{-13}, 2^{-15}]$ and kernel parameters $\gamma = [2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3, 2^5, 2^7, 2^9, 2^{11}, 2^{13}, 2^{15}]$ for each classification. For the linear SVM and the pyramid match kernel based SVM, the only parameter to select is the penalty parameter C, and their search space was set $C = [2^3, 2^1, 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}, 2^{-9}, 2^{-11}, 2^{-13}, 2^{-15}]$, with the parameters yielding the minimum five-fold cross-validation errors as their optimal SVM parameters.

To evaluate the representation ability of deep CNN features in inter-dataset classification, the common categories in every two datasets were picked out. The corresponding relations of common scene categories between every two datasets are given in Table 3.

To quantitatively evaluate the performances of deep features extracted from different pretrained CNN models, the classification result was evaluated by average overall classification accuracy from five runs on all scene categories. The overall classification accuracy is defined as the number of correctly predicted samples divided by the total number of testing samples. In different experiments, different datasets were used as the reference data for evaluation. Detailed information is given in Section 4.

Table 3 Corresponding relations of common scene categories between every two datasets

Class Index	Dataset-1	Dataset-2	Dataset-1	Dataset-3	Dataset-2	Dataset-3
1	Agricultural	Farmland	Buildings	High buildings	Commercial area	High buildings
2	Beach	Beach	Dense & medium density residential	Residential area	Forest	Dense forest
3	Buildings	Commercial area	Forest	Dense forest	Industrial area	Low buildings
4	Dense & medium density residential	Residential area	Harbor	Harbor	Meadow	Grassland
5	Forest	Forest	Overpass	Overpass	Railway station	Railway
6	Overpass	Viaduct	Storage tanks	Storage tanks	Residential area	Residential area
7	Parking lot	Parking lot			Viaduct	Overpass
8	River	River				

4 Results

To evaluate the deep CNN features in VHSR remote sensing image scene classification, the involved factors in scene classification were intensively compared, in which half of the images in the original datasets were used as training samples and the rest half were used as testing samples. Then, every two datasets were used as training and testing data in cross-classification to test the inter-dataset representation ability of CNN features and only the common categories were involved, as shown in Table 3. Take the cross-classification case of Dataset-1 and Dataset-2 as an example. There are eight common categories that both of the datasets share. If the images from the eight categories in Dasetset-1 are used as training data, the images from the eight categories in Dasetset-2 will then be used as testing data.

4.1 Comparison of involved factors

For AlexNet and Caffe, they have a similar architecture that contains three fully-connected layers and five convolutional layers. For VGG-16, it contains 13 convolutional layers and three fully-connected layers. For these three CNN architectures, their deep CNN features were extracted from the last convolutional layer (Conv5/Conv5_3), the first fully-connected layer (FC6), and the second fully-connected layer (FC7). However, for GoogLeNet, as it only has one fully-connected layer before output, the last convolutional layer [inception (5b)] and the last fully-connected layer (FC) were used to extract its deep CNN features. Note that the last fully-connected layer is not considered for all the models because it is actually a softmax layer that computes the scores for each defined class by specific tasks, and thus this layer does not contain so much general feature information as previous layers. Comparison results are shown in Fig. 10 and Table 4.

As shown in Fig. 10, the classification results are promising. To be specific, the best classification accuracies for Dataset-1 and Dataset-2 can exceed 94% and the best result for Dataset-3 can reach 90%, which indicates that the deep CNN features extracted from the pretrained CNN models are very good at capturing the essential characteristics of the remote sensing image scenes. For comparison of different kernels, the results obtained by the linear kernel are consistently better than those done by the RBF kernel. This is because the RBF kernel usually maps the samples from the original low dimensional feature space to a higher dimensional space, whereas the deep CNN features extracted from the convolutional or fully-connected layers already have very high dimensions, e.g., the Caffe model has a 4096-dimensional feature vector for the FC6 layer and a 43,264-dimensional feature vector for the Conv5 layer. Under such circumstances, the original feature space is high enough for discriminating different classes and does not need to be mapped into a much higher feature space. Therefore, the RBF kernel becomes unsuitable for such high dimensional feature vectors. In particular, for the high dimensional features from the convolutional layers, the higher the feature dimension is, the worse the RBF kernel performs, compared with the linear kernel, which can be corroborated by Fig. 10a, d, and g. Take VGG-16 and GoogLeNet as an example. For VGG-16, the feature vector from the last convolutional layer (Conv5_3) has a dimension of 25,088; for GoogLeNet, the feature dimension of the last convolutional layer [inception (5b)] reaches 50,176, one time higher than that from the VGG-16 architecture. When the convolutional layers are utilized to extract CNN features, the linear SVM can improve the classification accuracy of VGG-16 by around 10%, compared with the RBF based SVM, whereas the improvement for GoogLeNet is much more, reaching more than

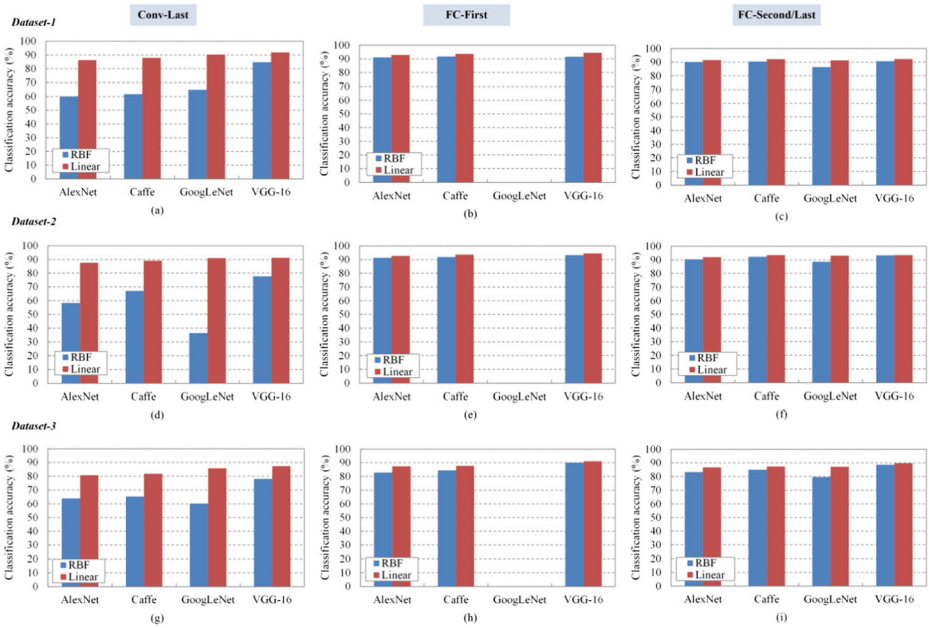

Fig. 10 Comparison of classification results with CNN features by classifiers of different kernels: using the last convolutional layer for classification of **a** Dataset-1, **d** Dataset-2, **g** Dataset-3; using the first fully-connected layer for classification of **b** Dataset-1, **e** Dataset-2, **h** Dataset-3; using the second/last fully-connected layer for classification of **c** Dataset-1, **f** Dataset-2, **i** Dataset-3

25%. This just verifies that the feature dimension does have impacts on the classification performance of the RBF kernel. Note that in Fig. 10d, the extremely low classification accuracy of GoogLeNet using the RBF based SVM may also result from the characteristics of the RBF kernel. In the following experiments, the linear SVM will be used for learning deep CNN features.

Table 4 Comparison of classification results using CNN features of different pretrained models and the SPM hand-crafted feature

	Models	Conv-Last(%)	FC-First(%)	FC-Second/Last(%)	Hand-crafted feature(%)
Dataset-1	AlexNet	86.30	92.90	91.62	\
	Caffe	87.98	93.66	92.23	\
	GoogLeNet	90.30	\	91.39	
	VGG-16	91.90	94.50	92.32	\
	SPM(baseline)	\	\		78.72
Dataset-2	AlexNet	87.67	92.84	92.09	\
	Caffe	89.15	93.72	93.56	\
	GoogLeNet	91.05	\	93.20	
	VGG-16	91.33	94.63	93.56	\
	SPM(baseline)	\	\		82.00
Dataset-3	AlexNet	80.75	87.34	86.72	\
	Caffe	81.79	87.63	87.31	\
	GoogLeNet	85.73	\	87.14	
	VGG-16	87.34	91.04	89.74	\
	SPM(baseline)	\	\		79.06

In Table 4, comparisons are made for the classification results obtained using features from different layers of pretrained CNN models. As can be seen in Table 4, the classification accuracies of different pretrained CNN models are quite similar, with GoogLeNet and VGG-16 slightly better than the other two CNN models in general, and the deep CNN features significantly outperform the SPM features in classification accuracy on all the three datasets, which indicates that the CNN features have relatively stronger feature representation abilities, compared with the hand-crafted features. The probable reason is that with such a brain-like feature learning mode, the feature representation learned from the CNN model is very close to the high-level semantic information of the whole image. Besides, by comparing the numbers of hidden layers, it can be observed that both GoogLeNet and VGG-16 have deeper architectures than the other two CNN models that only have eight hidden layers. With the increase of the number of hidden layers, the extracted features are deeper and become closer to high-level semantic information, which makes the feature vector of the scene image more descriptive. That is why AlexNet and Caffe have very close classification accuracies and why GoogLeNet and VGG-16 show relatively better classification results. These results just prove that the depth of the CNN architecture does play an important role in the feature representation ability of the extracted deep features. For each of the four pretrained CNN models, the features extracted from the fully-connected layers consistently excel those from the convolutional layers, which can be attributed to the fact that the fully-connected layers can usually learn more abstract semantic information than the convolutional layers. For the two fully-connected layers, the classification results from them both are more or less the same, but the features from the first fully-connected layers do slightly better in most cases.

4.2 Evaluation of inter-dataset feature representation ability

To evaluate the representation ability of CNN features in inter-dataset classification, each dataset was utilized as the training or testing samples, with the entire samples involved. Note that these three datasets are quite different in image size, spatial resolution, class patterns, and intra-class complexity. This experiment will test whether or not the CNN features are able to be learned and transferred to predict scene images from heterogeneous sources. As analyzed in the previous experiment, the first fully-connected layer of the pretrained CNN models will be applied to extract deep features, with the linear SVM as the classifier. Note that the GoogLeNet architecture only has one fully-connected layer, so this layer will be used to extract deep features. Table 5 gives the comparison results for cross-classification among three datasets using different pretrained CNN models. As shown in Table 5, the GoogLeNet and VGG-16 models do comparatively better than the other CNN models in classification accuracy, which

Table 5 Comparison of cross-classification results among three datasets using the CNN features of different pretrained models

Training- > Testing	AlexNet(%)	Caffe(%)	GoogLeNet(%)	VGG-16(%)
Dataset-1- > Dataset-2	67.92	74.94	62.53	70.96
Dataset-2- > Dataset-1	64.56	53.33	65.89	59.11
Dataset-1- > Dataset-3	84.99	84.11	87.03	90.09
Dataset-3- > Dataset-1	80.71	81.14	85.43	81.57
Dataset-2- > Dataset-3	54.59	51.02	61.10	65.56
Dataset-3- > Dataset-2	58.96	59.22	62.86	68.57

further confirms the positive effect of deep hidden layers on the classification performance. For cross-classification of Dataset-1 and Dataset-3, the classification accuracies using the CNN features can reach above 80%, however, for cross-classifications of Dataset-1 and Dataset-2, as well as Dataset-2 and Dataset-3, the classification performances using varied pretrained CNN models are not high and even the best accuracies only reach around 60%. The probable reason may be that Dataset-1 and Dataset-3 are almost identical in spatial resolution and their common categories are also similar in vision, whereas the spatial resolution of Dataset-2 is a little different from that of Dataset-1 and Dataset-3, which may lead to visual differences for some common scene categories under different spatial resolutions. These relatively low classification results also reveal that the feature representation capability of the deep features is limited and it needs to work with some conditions to meet with.

Besides, the classification accuracies of Dataset-3- > Dataset-1 are relatively worse than those of Dataset-1- > Dataset-3 for all the pretrained CNN models, which may result from the reason that compared with Dataset-1 that has an image size of 256 × 256, Dataset-3 has a larger image size (512 × 512) that is quite different from the input image size (224 × 224) of all the pretrained CNN models, and when Dataset-3 is used as the training set, many useful information will be lost after image scaling, largely affecting the recognition capability of the learned classification model. By comparing the cases of Dataset-1- > Dataset-2 and Dataset-2- > Dataset-1, as well as Dataset-2- > Dataset-3 and Dataset-3- > Dataset-2, it can be observed that for the CNN features, when two datasets are different in spatial resolution, the classification result using the lower spatial resolution dataset (e.g. Dataset-2) as the training set and the higher spatial resolution dataset (e.g., Dataset-1 and Dataset-3) as the testing set is generally worse than that using the higher spatial resolution dataset as the training set and the lower spatial resolution dataset as the testing set. Such a phenomenon may suffer from the fact that 1) the lower spatial resolution dataset may discard more detailed spatial information and thus cannot, to some extent, cover the feature details of the higher resolution dataset; 2) as input images need to be resized to fit the pretrained CNN models, the Dataset-2 with the largest image size will lose relatively more information than the other two datasets after down sampling.

To further verify the negative effect of down sampling on the classification performance, we conducted another experiment using Dataset-3. To perform the experiment, half of the dataset was randomly divided for five different runs and used as the training set, with the remaining half as the testing set. For the images in the training set, they were down sampled from 512 × 512 to 128 × 128 to generate a down sampled set. Then we used the original training set or the corresponding down sampled set as training samples and used the testing set as testing samples (without down sampling), so that it can be tested whether down sampling will bring down the classification performance. Here, we use the down sampled set and the dataset without down sampling as two different datasets of different image sizes. Table 6

Table 6 Comparison of classification results on Dataset-3 using the CNN features of different pretrained models with and without image down sampling

Methods	Without down sampling	With down sampling
AlexNet(%)	89.98	80.56
Caffe(%)	88.80	82.06
GoogLeNet(%)	90.42	83.48
VGG-16(%)	92.98	89.98

shows the comparison results. Note that in the table, "without down sampling" means using the original training set as training samples (512 × 512) and the testing set as testing samples (512 × 512), while "with down sampling" means using the down sampled training set as training samples (128 × 128) and the testing set as testing samples (512 × 512). As can be seen in Table 6, for all the compared CNN models, the classification accuracies using their extracted deep features decrease obviously after image down sampling of training samples. This further confirms that down sampling does have a negative effect on the scene classification performance and coincides with the analytical results of Table 5.

To make intensive comparisons, categorical classification accuracies are given in Fig. 11 with all the class indexes corresponding to the class names in Table 3. As shown in Fig. 11, most of the common scene categories can be correctly classified using the CNN features, except for a few cases, e.g., class 3 (buildings/commercial area) in the case of Dataset-1- > Dataset-2, classes 3 (commercial area/ buildings) and 6 (viaduct/overpass) in the case of Dataset-2- > Dataset-1, class 1 (commercial area/high buildings) in the case of Dataset-2- > Dataset-3, and classes 1 (high buildings/ commercial area) and 3 (low buildings/industrial area) in the case of Dataset-3- > Dataset-2. Figure 12 visually gives these wrongly classified categories among different datasets using the CNN features. It is apparent from Fig. 12 that the

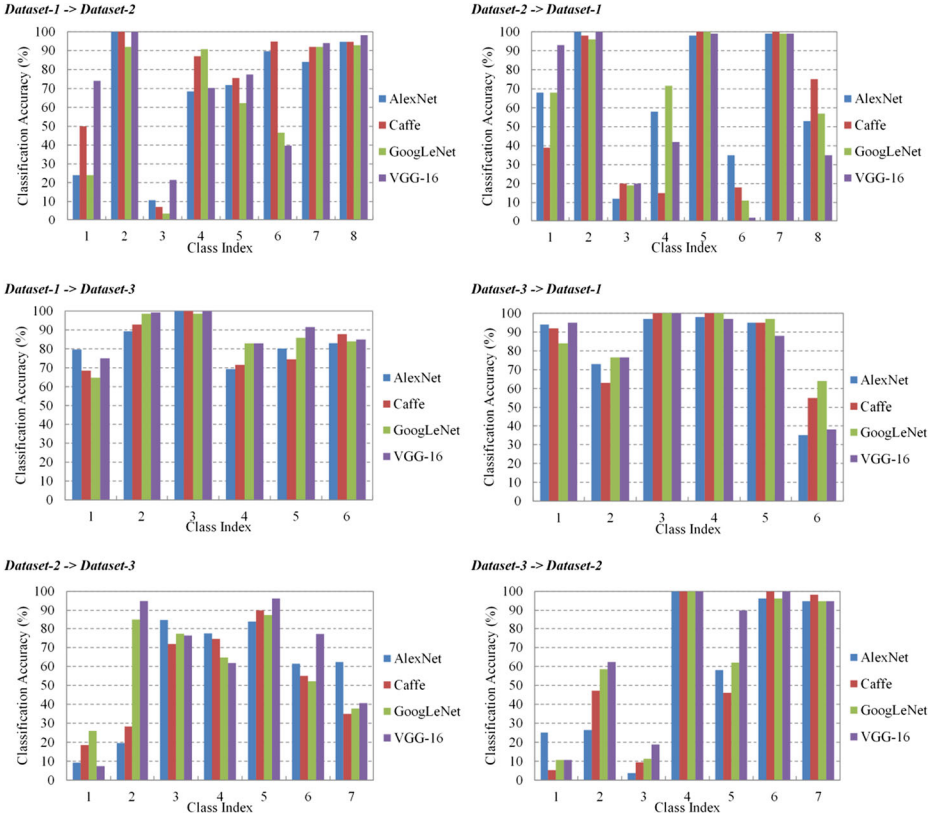

Fig. 11 Comparison of cross-classification results among three datasets for common categories: **a** Dataset-1 as training and Dataset-2 as testing; **b** Dataset-2 as training and Dataset-1 as testing; **c** Dataset-1 as training and Dataset-3 as testing; **d** Dataset-3 as training and Dataset-1 as testing; **e** Dataset-2 as training and Dataset-3 as testing; **f** Dataset-3 as training and Dataset-2 as testing

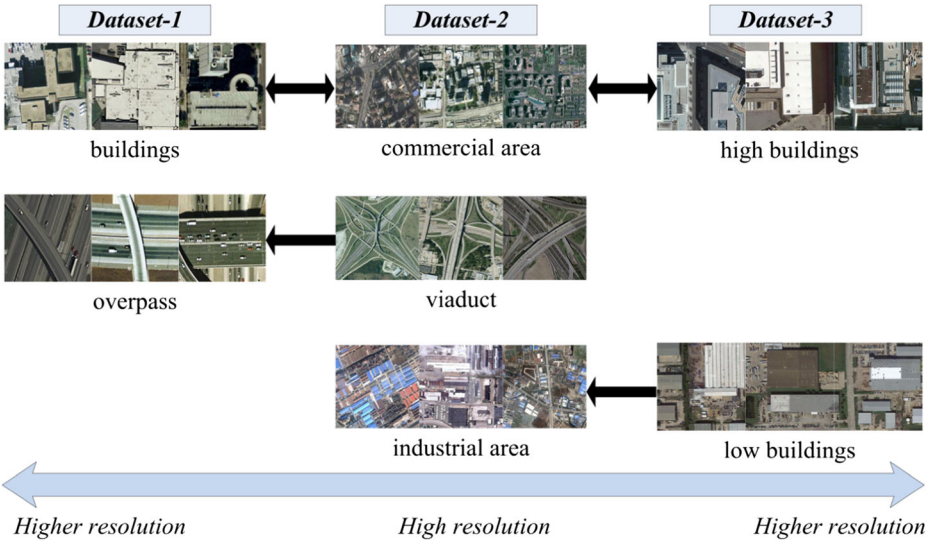

Fig. 12 Examples of wrongly classified categories among different datasets

low classification accuracies of these scene categories are likely to result from the big difference in spatial resolution, even though the common categories have the similar semantic information.

Table 7 shows the running time for feature extraction for a sample scene image using different pretrained CNN models. Here, the experimental environment is an Ubuntu 14.04 operating system with a 3.6 GHz Intel Core i7-4790 CPU and 8GB memory. It can be seen that both GoogLeNet and VGG-16 have slightly lower computing efficiency than other models and VGG-16 becomes relatively the most time-consuming one in feature extraction, but there does not exist much difference. The probable reason that affects the efficiency lies in the depth and complexity of the pretrained CNN model itself, as both VGG-16 and GoogLeNet are much deeper than the other two compared models in network architecture.

5 Discussion

From the above experimental results obtained using different kernels, it can be confirmed that the linear kernel is more suitable for high dimensional CNN features than the RBF kernel in SVM classification, which just coincide with the results of the previous work [17]. For all the compared models, the fully-connected layers have relatively better descriptive capabilities than the convolutional layers.

Table 7 Comparison of running time for feature extraction of a sample scene image from Dataset-1 using different pretrained CNN models

Models	AlexNet	Caffe	GoogLeNet	VGG-16
	0.1 s	0.09 s	0.26 s	0.84 s

For the analysis of the feature representation ability, the performance of the deep CNN features seems to be most sensitive to the variation of spatial resolution. Their representation ability in inter-dataset classification is also affected by the image size which, however, is a secondary factor. In the case that the spatial resolutions of the training dataset and the dataset for prediction are nearly the same, the CNN features can achieve a relatively high classification accuracy, effectively realizing the feature transfer and generalization among datasets from different sources. However, under the circumstance that the spatial resolutions of two datasets are different, such a feature representation ability will expose its weakness, especially when the image sizes of these two datasets are also much varied. These observations can serve as guidance for deep CNN features based VHSR remote sensing image scene classification in practical applications. To predict new scene images from heterogeneous sources, classification models using the deep features under different spatial resolutions can be constructed in advance. As long as the resolution information of the new image is known, we can choose the corresponding classification model for category prediction. Thanks to the powerful feature mining capability of deep CNN models, an opportunity is provided to avoid the problem that new classification models have to be retrained based on new classification tasks using new training samples.

In future applications, the scene classification of large satellite images with more than three spectral bands becomes a realistic problem. To meet the requirements of the pretrained CNN models, the principal component analysis (PCA) transform can be applied to reduce the spectral dimension. To realize the classification of large satellite scenes, the overlapping grid partition [50] or the superpixel segmentation [35] can firstly be applied to obtain the subimages to recognize, and then the pretrained CNN models can further be used to extract deep features.

6 Conclusions

Hitherto, few studies have looked into the strength and weakness of the deep CNN feature itself in practical inter-dataset scene classification of VHSR remote sensing images. There still leaves a question whether these deep CNN features have a powerful feature representation ability and are able to explore the essential attributes of scenes of the same kind but with different imaging conditions. This issue is of great significance since in practical applications, the scene image to recognize is not always from the same dataset as the training data. To deal with this problem, a comprehensive study is performed to discuss the characteristics of the deep CNN features and to investigate the feature representation ability in inter-dataset classification. Experimental results with four well-known pretrained CNN models on three different datasets revealed that 1) for remote sensing scene images of the same kind, the deep CNN features are sensitive to the changes of spatial resolutions, especially when there is an obvious resolution gap between training and prediction images, which indicates that such deep features are unable to explore the invariant semantic attributes of scenes from behind the remote sensing images under different resolutions; 2) the variation of image sizes between datasets has a negative impact on CNN feature representation but such an effect is minor, especially compared with the resolution factor; 3) provided that the spatial resolutions of the training and prediction scene images are close, the CNN features show satisfactory representation ability, under which circumstances the classification model learned from the deep CNN features of one dataset can be effectively applied to distinguish, with a high accuracy, the scene images from another different dataset; 4) among all the compared pretrained CNN models, the VGG-

16 and GoogLeNet models, as a feature extractor, can extract comparatively more powerful feature vectors in both intra- and inter-dataset scene classification tasks, which shows that it is helpful to extract deep CNN features using a relatively deeper network.

In the ongoing studies, further researches need to be conducted to investigate the effect of deeper pretrained CNN architectures on the performance of remote sensing image scene feature representation, such as ResNet. Besides, efforts should be made to improve the deep CNN features and make them resolution-invariant when dealing with inter-dataset scene classification problems.

Acknowledgments This work was supported in part by the Major Project of High Resolution Earth Observation System of China under Grant 03-Y20A04-9001-17/18 and in part by the National Natural Science Foundation of China under Grant 41701397.

References

1. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828. https://doi.org/10.1109/TPAMI.2013.50
2. Cai SS, Liu DS (2013) A comparison of object-based and contextual pixel-based classifications using high and medium spatial resolution images. Remote Sens Lett 4:998–1007. https://doi.org/10.1080/2150704X.2013.828180
3. Cao YH, Xu RF, Chen T (2015) Combining convolutional neural network and support vector machine for sentiment classification. Paper presented at the 4th National Conference on Social Media Processing, Guangzhou, China, November 16–17
4. Castelluccio M, Poggi G, Sansone C, Verdoliva L (2015) Land use classification in remote sensing images by convolutional neural networks. Available online: http://arxiv.org/abs/1508.00092. Accessed on 5 Nov 2016
5. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:1–27. https://doi.org/10.1145/1961189.1961199
6. Chen SZ, Tian YL (2015) Pyramid of spatial relatons for scene-level land use classification. IEEE Trans Geosci Remote Sens 53:1947–1957. https://doi.org/10.1109/TGRS.2014.2351395
7. Chen C, Zhang B, Su H, Li W, Wang L (2016) Land-use scene classification using multi-scale completed local binary patterns. SIViP 10:745–752. https://doi.org/10.1007/s11760-015-0804-2
8. Chen J, Song X, Nie L, Wang X, Zhang H, Chua TS (2016) Micro tells macro: predicting the popularity of micro-videos via a transductive model. Paper presented at the 2016 ACM Conference on Multimedia, Amsterdam, The Netherlands, October 15–19
9. Cheng G, Guo L, Zhao TY, Han JW, Li HH, Fang J (2013) Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int J Remote Sens 34:45–59. https://doi.org/10.1080/01431161.2012.705443
10. Cheng G, Han J, Guo L, Liu Z, Bu S, Ren J (2015) Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Trans Geosci Remote Sens 53:4238–4249. https://doi.org/10.1109/TGRS.2015.2393857
11. Cheriyadat AM (2014) Unsupervised feature learning for aerial scene classification. IEEE Trans Geosci Remote Sens 52:439–451. https://doi.org/10.1109/TGRS.2013.2241444
12. Csurka G, Dance CR, Fan LX, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. Paper presented at the 2004 ECCV International Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, May 11–14
13. Dai D, Yang W (2011) Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geosci Remote Sens Lett 8:173–176. https://doi.org/10.1109/LGRS.2010.2055033
14. Duro DC, Franklin SE, Dube MG (2012) A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens Environ 118:259–272. https://doi.org/10.1016/j.rse.2011.11.020

15. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507. https://doi.org/10.1126/science.1127647
16. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machine. IEEE Trans Neural Netw 13:415–425. https://doi.org/10.1109/72.991427
17. Hu F, Xia GS, Hu J, Zhang L (2015) Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens 7:14680–14707. https://doi.org/10.3390/rs71114680
18. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. Available online: http://arxiv.org/abs/1408.5093. Accessed on 26 Sept 2016
19. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Paper presented at the 26th Annual Conference on Neural Information Processing Systems, Harrahs and Harveys, Lake Tahoe, USA, December 3–8
20. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, June 17–22
21. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1:541–551. https://doi.org/10.1162/neco.1989.1.4.541
22. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324
23. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/Nature14539
24. Luus FPS, Salmon BP, van den Bergh F, Maharaj BTJ (2015) Multiview deep learning for land-use classification. IEEE Geosci Remote Sens Lett 12:2448–2452. https://doi.org/10.1109/LGRS.2015.2483680
25. Marmanis D, Datcu M, Esch T, Stilla U (2016) Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geosci Remote Sens Lett 13:105–109. https://doi.org/10.1109/LGRS.2015.2499239
26. Mekhalfi ML, Melgani F, Bazi Y, Alajlan N (2015) Land-use classification with compressive sensing multifeature fusion. IEEE Geosci Remote Sens Lett 2155–2159:12. https://doi.org/10.1109/LGRS.2015.2453130
27. Muhling M, Korfhage N, Muller E, Otto C, Springstein M, Langelage T, Veith U, Ewerth R, Freisleben B (2017) Deep learning for content-based video retrieval in film and television production. Multimed Tools Appl 76:22169–22194. https://doi.org/10.1007/s11042-017-4962-9
28. Nogueira K, Penatti OAB, dos Santos JA (2017) Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn 61:539–556. https://doi.org/10.1016/j.patcog.2016.07.001
29. Oommen T, Misra D, Twarakavi NKC, Prakash A, Sahoo B, Bandopadhyay S (2008) An objective analysis of support vector machine based classification for remote sensing. Math Geosci 40:409–424. https://doi.org/10.1007/s11004-008-9156-6
30. Penatti OAB, Nogueira K, dos Santos JA (2015) Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshop, Boston, MA, USA, June 7–12
31. Qi K, Wu H, Shen C, Gong J (2015) Land-use scene classification in high-resolution remote sensing images using improved correlatons. IEEE Geosci Remote Sens Lett 12:2403–2407. https://doi.org/10.1109/LGRS.2015.2478966
32. Qu T, Zhang QY, Sun SL (2017) Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks. Multimed Tools Appl 76:21651–21663. https://doi.org/10.1007/s11042-016-4043-5
33. Salberg AB (2015) Detection of seals in remote sensing images using features extracted from deep convolutional neural networks. Paper presented at the IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, July 26–31
34. Shahriari M, Bergevin R (2017) Land-use scene classification: a comparative study on bag of visual word framework. Multimed Tools Appl 76:23059. https://doi.org/10.1007/s11042-016-4316-z
35. Shao W, Yang W, Xia GS (2013) Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification. Int J Remote Sens 34:8588–8602. https://doi.org/10.1080/01431161.2013.845925
36. Sheng GF, Yang W, Xu T, Sun H (2012) High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int J Remote Sens 33:2395–2412. https://doi.org/10.1080/01431161.2011.608740

37. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Available online: http://arxiv.org/abs/1409.1556. Accessed on 26 Sept 2016
38. Song X, Feng F, Liu J, Li Z, Nie L, Ma J (2017) NeuroStylist: neural compatibility modeling for clothing matching. Paper presented at the 2017 ACM Conference on Multimedia, Mountain View, CA, USA, October 23–27, 2017
39. Sridharan H, Cheriyadat A (2015) Bag of lines (BoL) for improved aerial scene representation. IEEE Geosci Remote Sens Lett 12:676–680. https://doi.org/10.1109/LGRS.2014.2357392
40. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 7-12
41. Wang Q, Lin J, Yuan Y (2016) Salient band selection for hyperspectral image classification via manifold ranking. IEEE Trans Neural Netw Learn Syst 27:1279–1289. https://doi.org/10.1109/TNNLS.2015.2477537
42. Weng Q, Mao Z, Lin J, Guo W (2017) Land-use classification via extreme learning classifier based on deep convolutional features. IEEE Geosci Remote Sens Lett 14:704–708. https://doi.org/10.1109/LGRS.2017.2672643
43. Whiteside TG, Boggs GS, Maier SW (2011) Comparing object-based and pixel-based classifications for mapping savannas. Int J Appl Earth Obs Geoinf 13:884–893. https://doi.org/10.1016/j.jag.2011.06.008
44. Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. Paper presented at the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November
45. Yu X, Wu X, Luo C, Ren P (2017) Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. GISci Remote Sens 54:741–758. https://doi.org/10.1080/15481603.2017.1323377
46. Zhao B, Zhong YF, Zhang LP (2013) Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. Remote Sens Lett 4:1204–1213. https://doi.org/10.1109/TPAMI.2007.70716
47. Zhao LJ, Tang P, Huo LZ (2014) A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification. Int J Remote Sens 35:2296–2310. https://doi.org/10.1080/01431161.2014.890762
48. Zhao LJ, Tang P, Huo LZ (2014) Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. IEEE J Sel Top Appl Earth Obs Remote Sens 7:4620–4631. https://doi.org/10.1109/JSTARS.2014.2339842
49. Zhao LJ, Tang P, Huo LZ (2016) Feature significance based multibag-of-visual-words model for remote sensing image scene classification. J Appl Remote Sens 10:035004. https://doi.org/10.1117/1.JRS.10.035004
50. Zhong YF, Zhu QQ, Zhang LP (2015) Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Trans Geosci Remote Sens 53:6207–6222. https://doi.org/10.1109/TGRS.2015.2435801

Lijun Zhao was born in Luoyang City, China, in 1986. He received the B.S. degree in geographical information system from Henan University, Kaifeng, China, in 2005 and the M.E. degree in environmental science from Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, China, in 2012. In 2015, he

received the Ph.D. degree in signal and information processing at Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. Currently, he is an Assistant Researcher with the Remote Sensing Image Processing Laboratory, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. His research interests include high-resolution remote sensing image understanding, image retrieval, and machine learning. Dr. Zhao is a Reviewer of the IEEE Transactions on Geoscience and Remote Sensing and the IEEE Geoscience and Remote Sensing Letters.

Wei Zhang was born in Yuncheng city of Shanxi province, China, in 1992. He received the B.E degree in environmental engineering from Dalian University, Dalian, China, in 2014. Later, in 2017, he received his M.E. degree in electronics and communication engineering from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. He is currently pursuing the Ph.D. degree of signal and information processing also at the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. His research interests include land-cover identification and classification of remote sensing images and machine learning.

Ping Tang received the B.S. degree in mathematics from Ningxia University, Yinchuan, China, in 1986, and the M.S. and Ph.D. degrees both in mathematics from Beijing Normal University, Beijing, China, in 1993 and 1996, respectively. In 1998, she was a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. Currently, she is serving as a Team Leader and leading a project of multispectral imagery radiometric and geometric correction for large volume of image data at global scale for higher resolution global land-cover mapping. Her expertise lies in using mathematical theories to develop algorithms related to image processing and analysis. She has significant experience in managing and designing software systems for satellite image processing and applications.