Deep Learning for Feature Extraction in Remote Sensing: A Case-study of Aerial Scene Classification

Biserka Petrovska, Eftim Zdravevski,
Petre Lameski, Roberto Corizzo,
Ivan Štajduhar and Jonatan Lerga

Computer vision

- Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world.

- Object classification, identification, verification, detection, segmentation

- High-level problems where we have seen success with computer vision: Optical character recognition (OCR), 3D model building, Medical imaging, Automotive safety, Fingerprint recognition and biometrics

Remote sensing (aerial scene) classification

- Aerial scene classification is process in which semantic label is assigned to images collected from remote locations

- Existence of several RS images datasets collected from satellites, aerial systems, and unmanned aerial vehicles (UAV)

- Military, surveillance and security, environment monitoring, detection of geospatial objects

Scene classification methods (1)

- methods that utilize low-level image features: spectral, textural, structural descriptors
- Scale Invariant Feature Transform (SIFT), Gabor texture features, color histograms, Grey Level Co-Occurrence Matrix (GLCM)
- mid-level visual representation methods: represent scenes with statistical representation of high-degree get from the extracted local image features

Scene classification methods (2)

- bag-of-visual-words (BoVW), sparse coding method, Principal Component Analysis (PCA), the Improved Fisher Vector (IFK), Vectors of Locally Aggregated Tensors (VLAT)

- methods using high-level image features: image classification, object recognition and image retrieval

- high-level methods can obtain more abstract and discriminative semantic representations

Scene classification methods (3)

- convolutional neural networks (CNNs)
- Feature extraction with convolutional neural networks (CNNs), pre-trained on massive data sets
- Fine-tuning of the weights of a pre-trained CNN
- optimize CNN from scratch

Contributions of the research

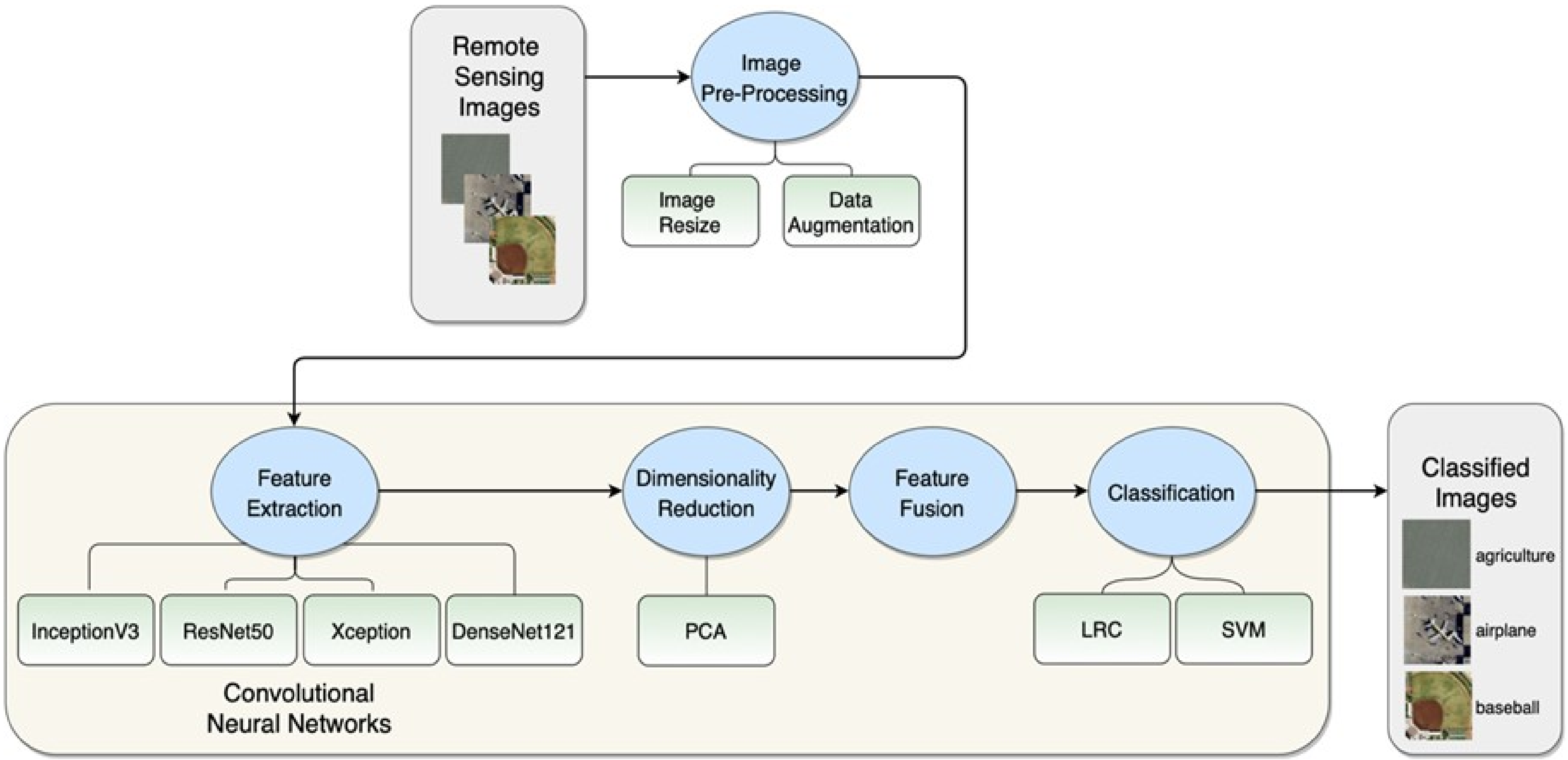
- CNN average pooling layers and convolutional layers are combined to generate image scene features

- InceptionV3, ResNet50, Xception and DenseNet121 for image features extraction

- Dimensionality reduction of the dense CNN activations using the PCA, feature fusion and evaluation based on Linear SVM and LRC

- Comparison to the existing methods on two publicly available remote sensing data sets

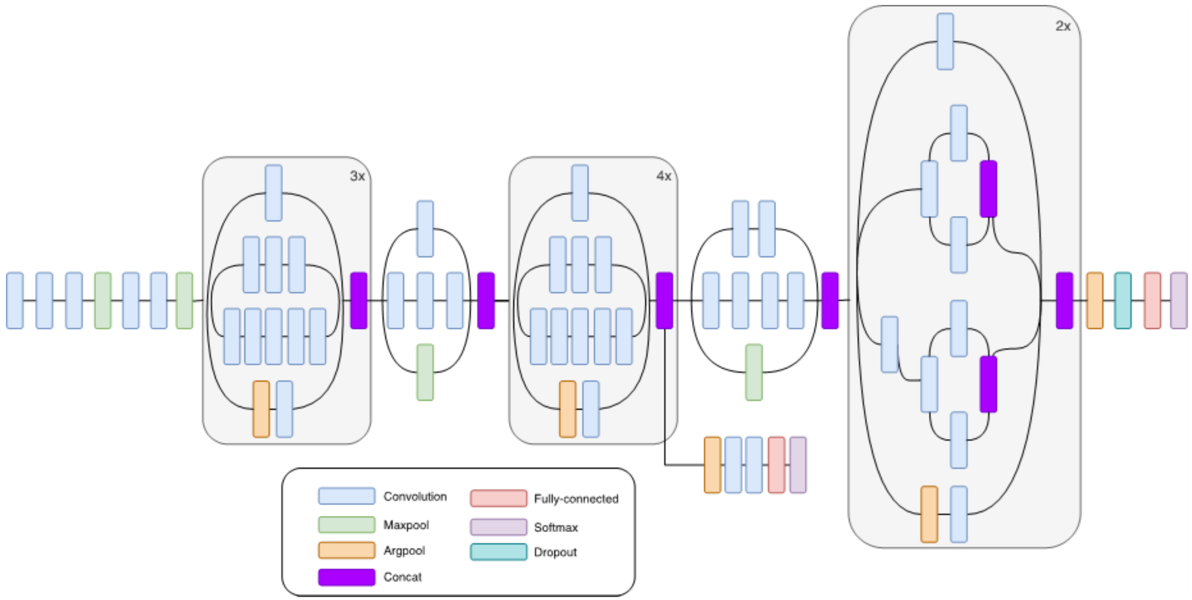
Workflow of the proposed method



CNNs for feature extraction (1)

- InceptionV3 - “Inception module”, contains 1x1, 3x3 and 5x5 convolutional layers and processes its input in a parallel workflow

- InceptionV3- can increase its depth and width without causing computational strain

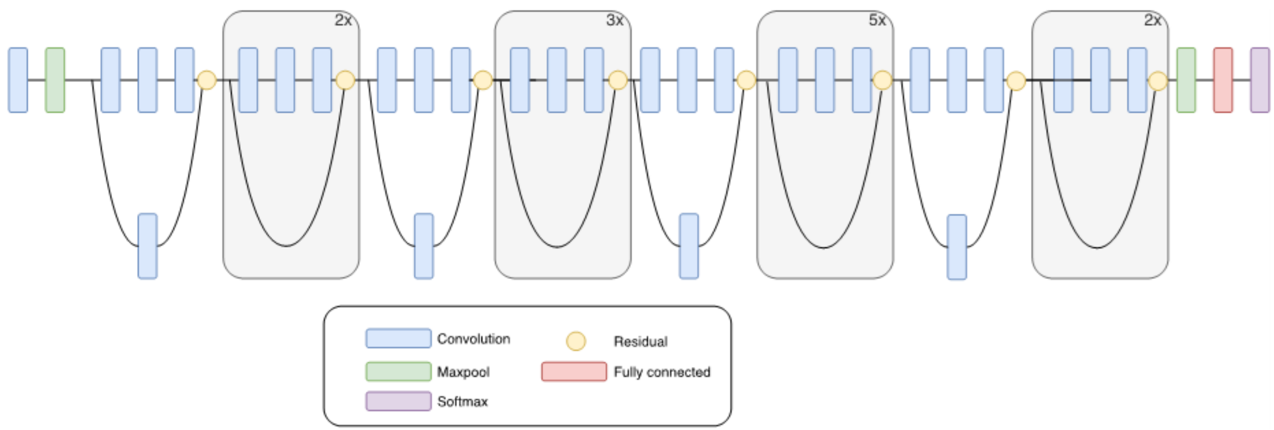


Schematic drawing of the InceptionV3 CNN

CNNs for feature extraction (2)

- ResNet50 – "Residual module" has a shortcut between the input and the output

- ResNet50 solves the problem of vanishing gradient with an application of residual module

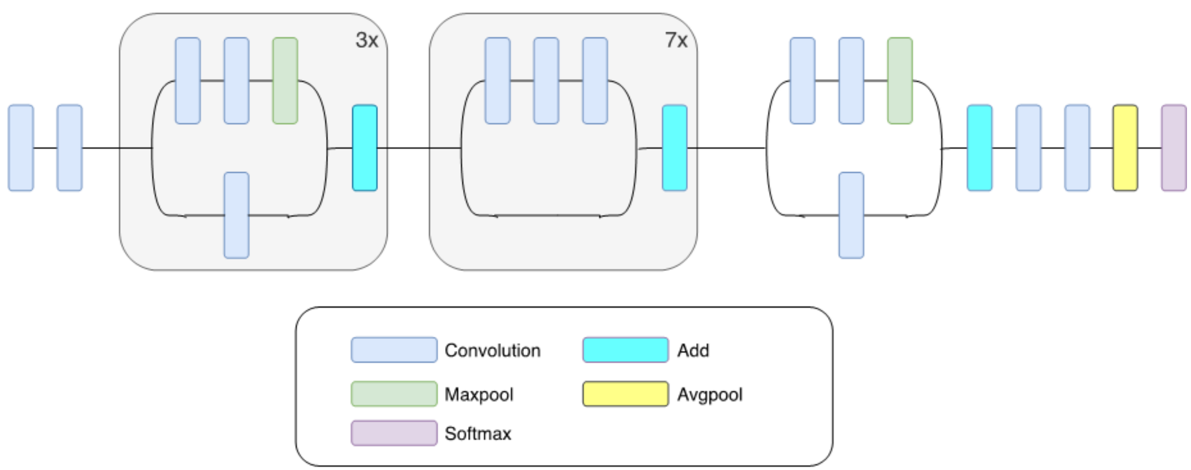


Schematic drawing of the ResNet CNN

CNNs for feature extraction (3)

- Xception –an assemble of depth-wise separable convolutional layers with residual connections

- Convolutions are performed in two steps: a depth-wise convolutional layer and a pointwise convolutional layer

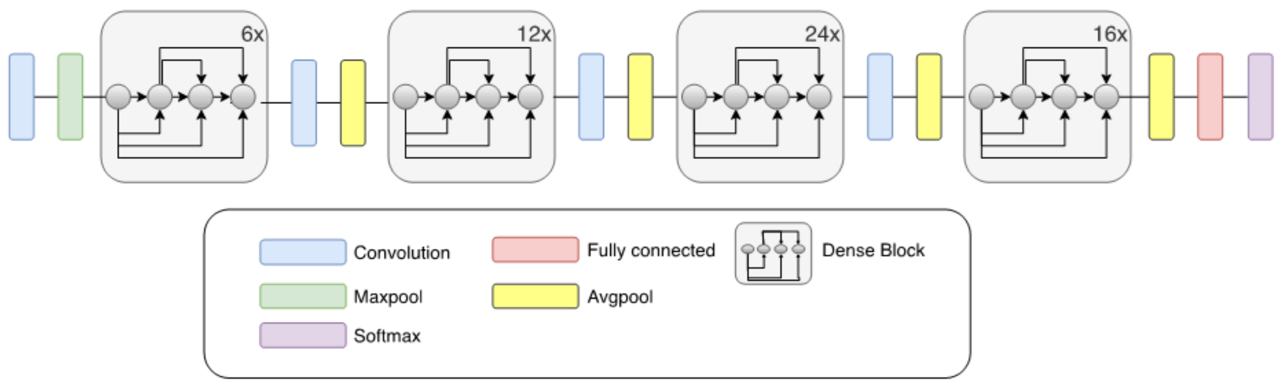


Schematic drawing of the Xception CNN

CNNs for feature extraction (4)

- DenseNet121 - each layer receives inputs from all previous layers, and it connects its outputs to every layer ahead

- **The number of parameters is decreased, the network is not prone to overfitting**



Schematic drawing of the DenseNet CNN

Principal Component Analysis

1. Calculate the covariance matrix X of input data points with dimensions m*n

2. Eigen vectors and corresponding eigen values should be calculated next

3. Order the eigen vectors according to their eigen, such that they are decreasing

4. New reduced k dimensions will be the first k eigen vectors

5. Transform the original n dimensional data points into k dimensions

Data sets (1)

- The UC-Merced data set has 2100 aerial scene images in 21 classes, their dimensions are **256x256 pixels**

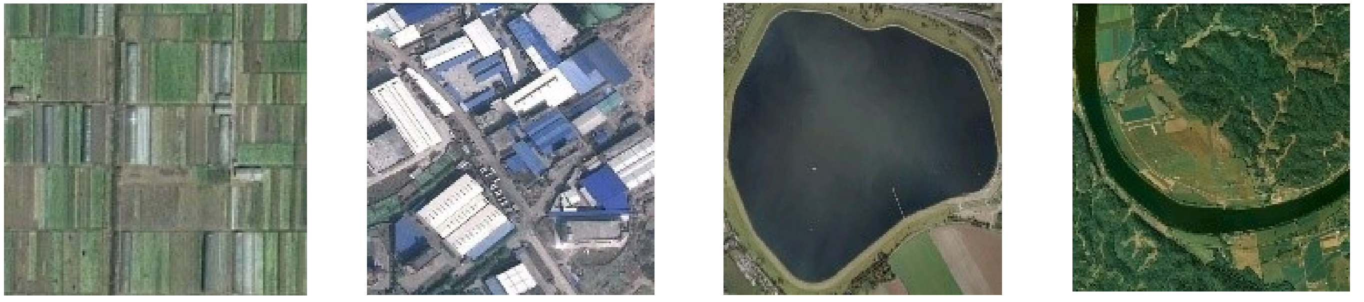
- The original images were downloaded from the United States Geological Survey (USGS) National Map

Some images of different classes from UC Merced dataset

Data sets (2)

- The WHU-RS data set is collected from Google Earth imagery

- There are 1005 images assigned to 19 classes with dimensions 600×600 pixels.



Some images of different classes from WHU-RS dataset

Experimental setup – 1^{st} simulation scenario (1)

- Feature extraction from three different CNNs' layers: average pooling and convolutional layers

- Image re-sizing, pre-processing, data augmentation on training images, no stratification

- Training/test data set ratio was 80% vs 20% for the UC-Merced data set, and 60% vs 40% for the WHU-RS data set

- After the features were extracted, a linear classifier was trained (LRC or SVM)

Experimental setup – 2^{nd} simulation scenario (1)

- the features were extracted from two different layers of two different CNNs

- moderate data augmentation on the training dataset, no stratification, image re-sizing and pre-processing

- before the feature fusion (concatenation), the PCA transformation is performed on features extracted from the convolutional layer

Experimental setup – 2nd simulation scenario (2)

- L2 normalization, feature concatenation, SVM classification

- Training/ test data set ratio was 80%/20% and 50%/50% for the UC-Merced data set, and 60%/40% and 40%/60% for the WHU-RS data set.

- Keras, TensorFlow, Nvidia GeForce GTX 1080 Ti with 11 GB of memory with CUDA v9.0

Evaluation metrics

- Classification accuracy is the ratio between the number of properly classified test images and the total number of test images

- The confusion matrix is a graphical display (table) of the classification accuracy on each class of the dataset

- In a normalized confusion matrix, item xij is the percentage of images that are classified as they belonged to i-th class, but their real class is j-th

Results - classification founded on extracted features from different CNN layers (1)

Method	LRC	SVM
ResNet50		
avg pooling	96.19	95.71
last conv layer	95.71	97.38
bn4f_branch2c	94.52	93.57
InceptionV3		
avg pooling	96.67	95
mixed_10	95.48	95.71
mixed_8	98.10	98.33
Xception		
avg pooling	93.57	94.76
block14_sepconv2_act	93.81	94.29
block14_sepconv1_act	96.43	95.71
DenseNet121		
avg pooling	95.48	93.81
conv5_block16_concat	96.67	94.05
conv4_block24_concat	97.14	95.24

Classification accuracy (OA (%)) of linear classification with LRC and SVM using features extracted from different layers with 80% of UC-Merced data set as training set

Results - classification founded on extracted features from different CNN layers (2)

Method	LRC	SVM
ResNet50		
avg pooling	98.01	97.01
last conv layer	98.01	97.76
bn4f_branch2c	95.52	96.02
InceptionV3		
avg pooling	95.78	95.02
mixed_10	94.53	95.52
mixed_8	97.26	97.26
Xception		
avg pooling	93.28	93.53
block14_sepconv2_act	94.28	94.53
block14_sepconv1_act	95.27	95.52
DenseNet121		
avg pooling	96.52	95.27
conv5_block16_concat	96.27	95.52
conv4_block24_concat	96.27	96.27

Classification accuracy (OA (%)) of linear classification with LRC and SVM of features extracted from different layers with 60% of WHU-RS data set as training set

Results - classification founded on extracted features from different CNN layers (3)

- the average pooling layer is a replacement for the fully connected layers and it gives features which represent the spatial dependencies between object parts, the whole object

- convolutional layers give features which represent mid-level information, object parts

- the best accuracies are obtained with mixed_8 layer for the InceptionV3, block14_sepconv1_act layer for the Xception, and conv4_block24_concat layer for the DenseNet121.

Results - classification based on features fusion with PCA transformation (1)

- PCA decomposition with 2010 components

- Linear SVM with grid search to select the value for C from the set of values: 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, and 100000

- The maximum iterations were 5000 and using 3-fold standard cross-validation using the training subset

Results - classification based on features fusion with PCA transformation (2)

Method	80% of UCM data set as training set	50% of UCM data set as training set
ResNet50 last conv layer (PCA) + InceptionV3 avg pooling	97.14	97.33
ResNet50 last conv layer (PCA) + Xception avg pooling	97.62	97.43
DenseNet121 conv5_block16_concat (PCA) + Xception avg pooling	97.86	96.67
DenseNet121 conv4_block24_concat (PCA) + Xception avg pooling	97.86	96.57
InceptionV3 mixed_10 (PCA) + ResNet50 avg pooling	97.62	96.57
InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling	98.33	97.43
InceptionV3 mixed_10 (PCA) + Xception avg pooling	95.95	95.14
InceptionV3 mixed_8 (PCA) + Xception avg pooling	98.57	97.62
DenseNet121 conv5_block16_concat (PCA) + ResNet50 avg pooling	97.14	96.67
DenseNet121 conv4_block24_concat (PCA) + ResNet50 avg pooling	96.9	95.24
Xception block14_sepconv2_act (PCA) + DenseNet121 avg pooling	96.67	96.48
Xception block14_sepconv1_act (PCA) + DenseNet121 avg pooling	98.57	96.29

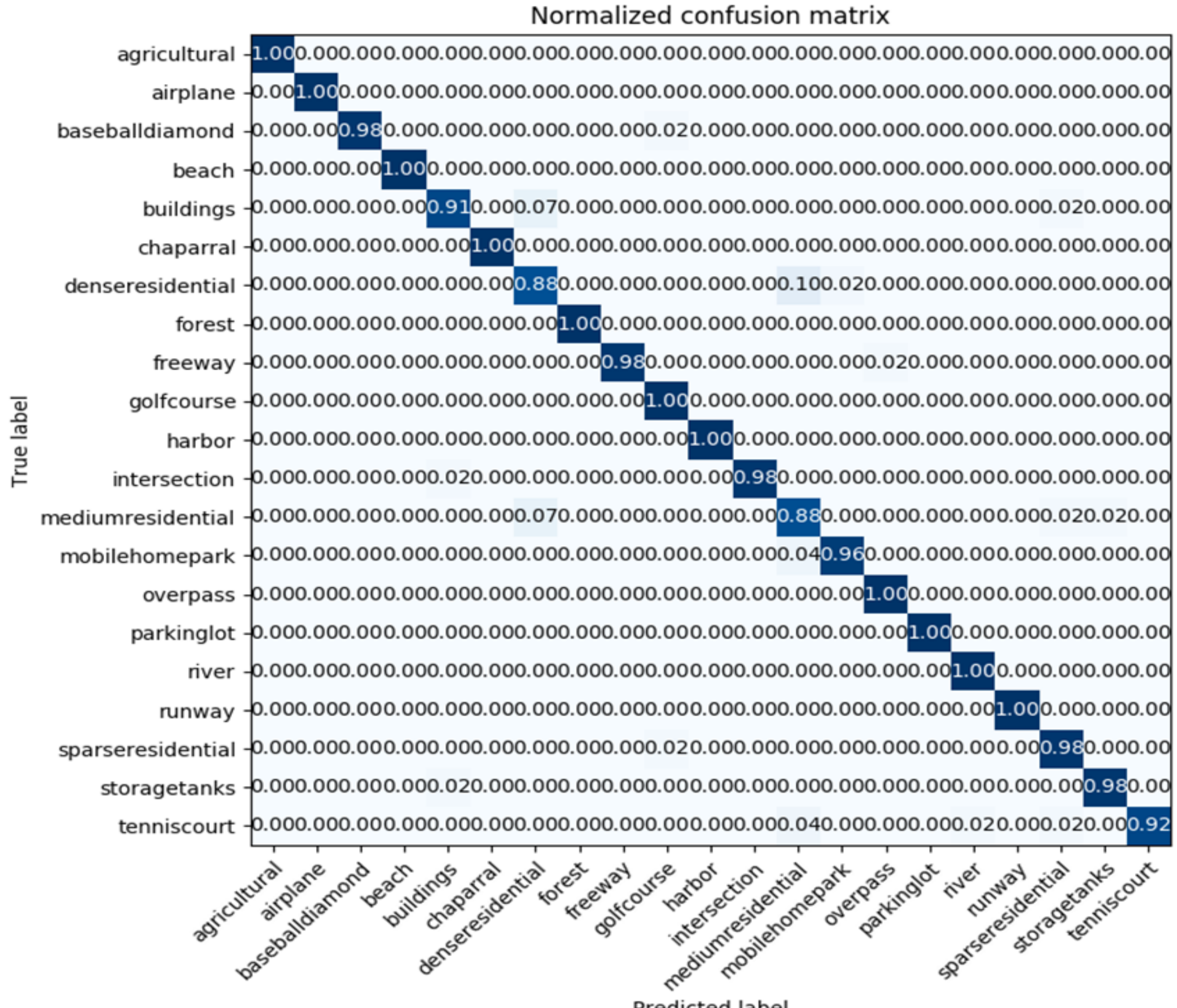
Classification accuracy (OA (%)) of linear classification of fused features with PCA transformation with 80% and 50% of UC-Merced data set as training set

Results - classification based on features fusion with PCA transformation (3)

Method	80% of UCM data set as training set	50% of UCM data set as training set
CaffeNet [45]	95.02 ± 0.81	93.98 ± 0.67
GoogLeNet [45]	94.31 ± 0.89	92.70 ± 0.60
VGG-16 [45]	95.21 ± 1.20	94.14 ± 0.69
SRSCNN [56]	95.57	/
CNN-ELM [57]	95.62	/
salM^{3}LBP-CLM [58]	95.75 ± 0.80	94.21 ± 0.75
TEX-Net-LF [59]	96.62 ± 0.49	95.89 ± 0.37
LGFBOVW [60]	96.88 ± 1.32	/
Fine-tuned GoogLeNet [61]	97.10	/
Fusion by addition [62]	97.42 ± 1.79	/
CCP-net [63]	97.52 ± 0.97	/
Two-stream Fusion [64]	98.02 ± 1.03	96.97 ± 0.75
DSFATN [65]	98.25	/
Deep CNN Transfer [38]	98.49	/
InceptionV3 mixed_8 (PCA) + Xception avg pooling (Ours)	**98.57**	**97.62**
GCFs+LOFs [66]	99 ± 0.35	97.37 ± 0.44
Inception-v3-CapsNet [55]	99.05 ± 0.24	97.59 ± 0.16

Classification accuracy (OA (%) and SD) of the examined method and the reference methods with 80% and 50% of UC-Merced data set as training set

Results - classification based on features fusion with PCA transformation (4)



Confusion matrix of the examined method with 50% of UC-Merced data set as training set for InceptionV3 mixed_8 (PCA) + Xception avg pooling

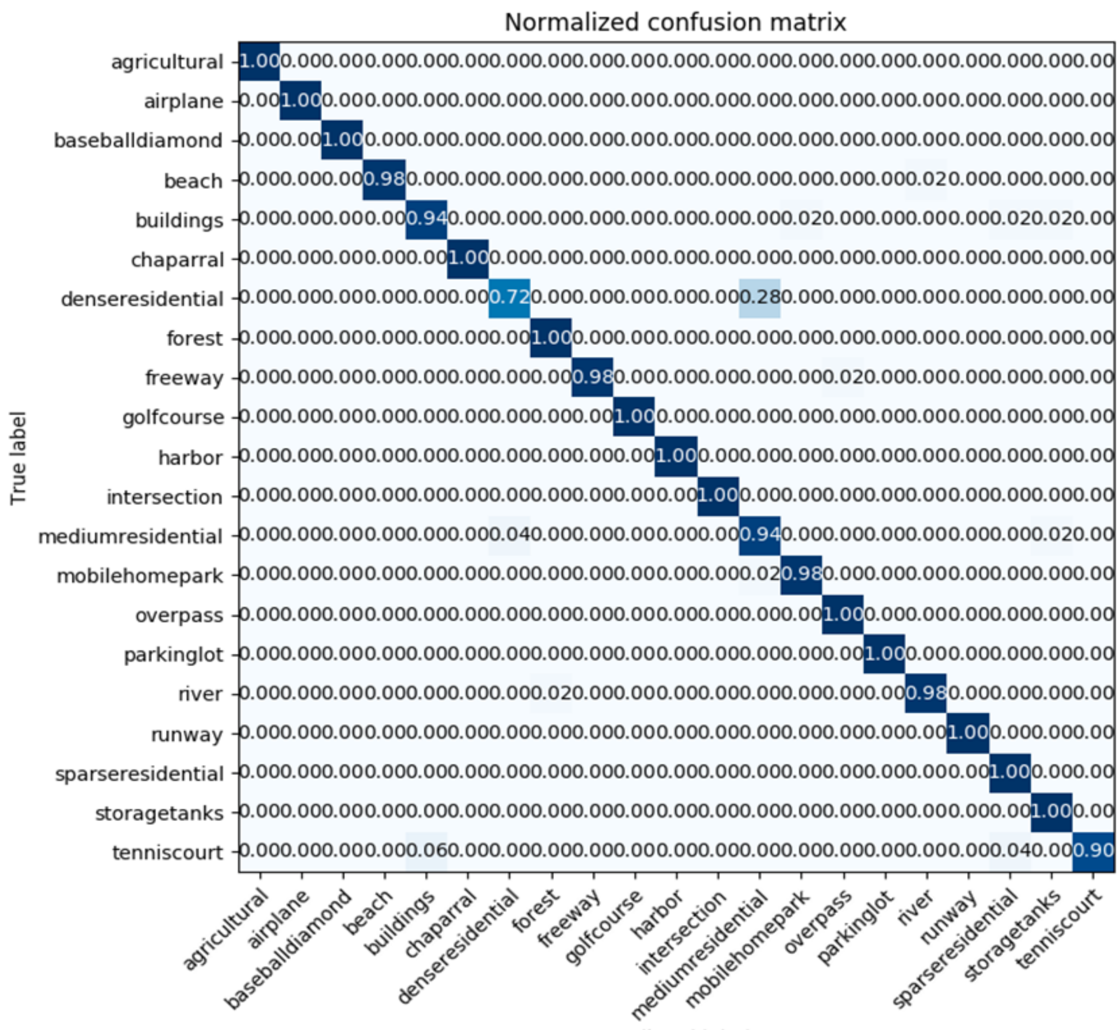
Results - classification based on features fusion with PCA transformation (5)

- Categories ‘dense residential’ and ‘medium residential’ achieved an accuracy of 88%

- GCFs+LOFs with ‘dense residential’ accuracy of 74%, and the Inception-v3-CapsNet with ‘dense residential’ accuracy of 80%

Class representatives of the UC Merced LandUse data set: (a) dense residential; (b) dense residential; (c) medium residential; (d) medium residential

Results - classification based on features fusion with PCA transformation (6)



Confusion matrix of the examined method with 50% of UC-Merced data set as training set for InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling

Results - classification based on features fusion with PCA transformation (7)

Method	60% of WHU-RS data set as training set	40% of WHU-RS data set as training set
ResNet50 last conv layer (PCA) + InceptionV3 avg pooling	98.26	95.02
ResNet50 last conv layer (PCA) + Xception avg pooling	97.62	96.52
DenseNet121 conv5_block16_concat (PCA) + Xception avg pooling	97.01	95.69
DenseNet121 conv4_block24_concat (PCA) + Xception avg pooling	97.76	96.68
InceptionV3 mixed_10 (PCA) + ResNet50 avg pooling	96.27	95.85
InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling	98.01	98.67
InceptionV3 mixed_10 (PCA) + Xception avg pooling	96.77	96.02
InceptionV3 mixed_8 (PCA) + Xception avg pooling	98.01	96.35
DenseNet121 conv5_block16_concat (PCA) + ResNet50 avg pooling	98.76	98.34
DenseNet121 conv4_block24_concat (PCA) + ResNet50 avg pooling	96.77	96.52
Xception block14_sepconv2_act (PCA) + DenseNet121 avg pooling	97.51	96.35
Xception block14_sepconv1_act (PCA) + DenseNet121 avg pooling	97.76	96.52
DenseNet121 conv5_block16_concat (PCA) + InceptionV3 avg pooling	96.27	97.51
DenseNet121 conv4_block24_concat (PCA) + InceptionV3 avg pooling	98.01	97.18

Classification accuracy (OA (%)) of linear classification of fused features with PCA transformation with 60% and 40% of WHU-RS data set as training set

Results - classification based on features fusion with PCA transformation (8)

- To check the reliability of results, all cases where the largest OA is obtained are repeated ten times on testing sets

- our proposed method for a training ratio of 40% outperforms all the other cutting-edge classification methods

Method	60% of WHU-RS data set as a training set	40% of WHU-RS data set as a training set
InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling	98.13 ± 0.51	97.84 ± 0.53
DenseNet121 conv5_block16_concat (PCA) + ResNet50 avg pooling	98.01 ± 0.68	98.26 ± 0.40

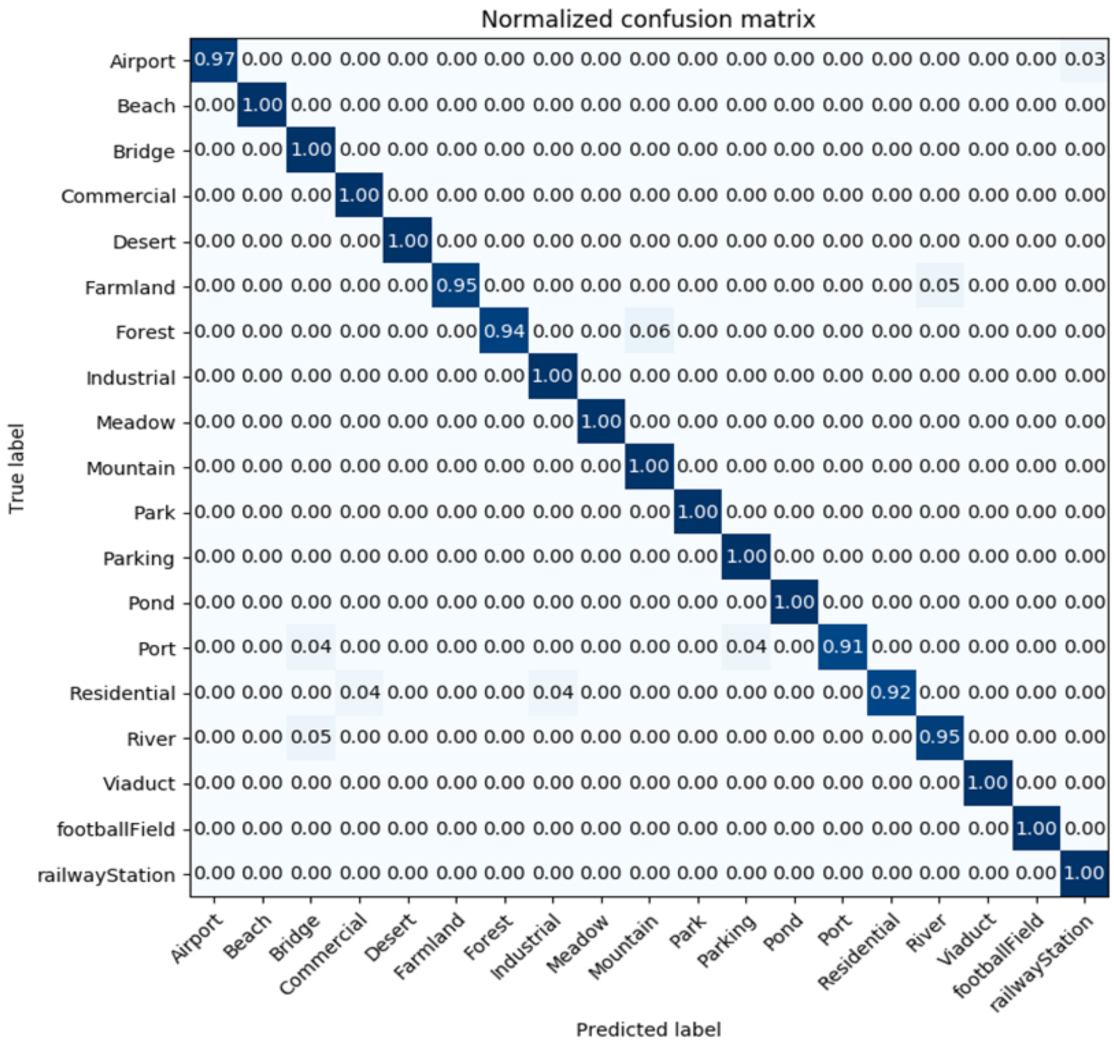
Classification accuracy (OA (%) and SD) of the examined method with 60% and 40% of the WHU-RS data set as a training set.

Results - classification based on features fusion with PCA transformation (9)

Method	60% of WHU-RS data set as training set	40% of WHU-RS data set as training set
Bag of SIFT [30]	85.52 ± 1.23	/
MS-CLBP + BoVW [67]	89.29 ± 1.30	/
GoogLeNet [45]	94.71 ± 1.33	93.12 ± 0.82
VGG-VD-16 [45]	96.05 ± 0.91	95.44 ± 0.60
CaffeNet [45]	96.24 ± 0.56	95.11 ± 1.20
salM^{3}LBP-CLM [58]	96.38 ± 0.82	95.35 ± 0.76
TEX-Net-LF [59]	96.62 ± 0.49	95.89 ± 0.37
InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling (Ours)	**98.13 ± 0.51**	/
DCA by addition [62]	98.70 ± 0.22	97.61 ± 0.36
Fusion with saliency detection [64]	98.92 ± 0.52	98.23 ± 0.56
DenseNet121 conv5_block16_concat (PCA) + ResNet50 avg pooling (Ours)	/	**98.26 ± 0.40**

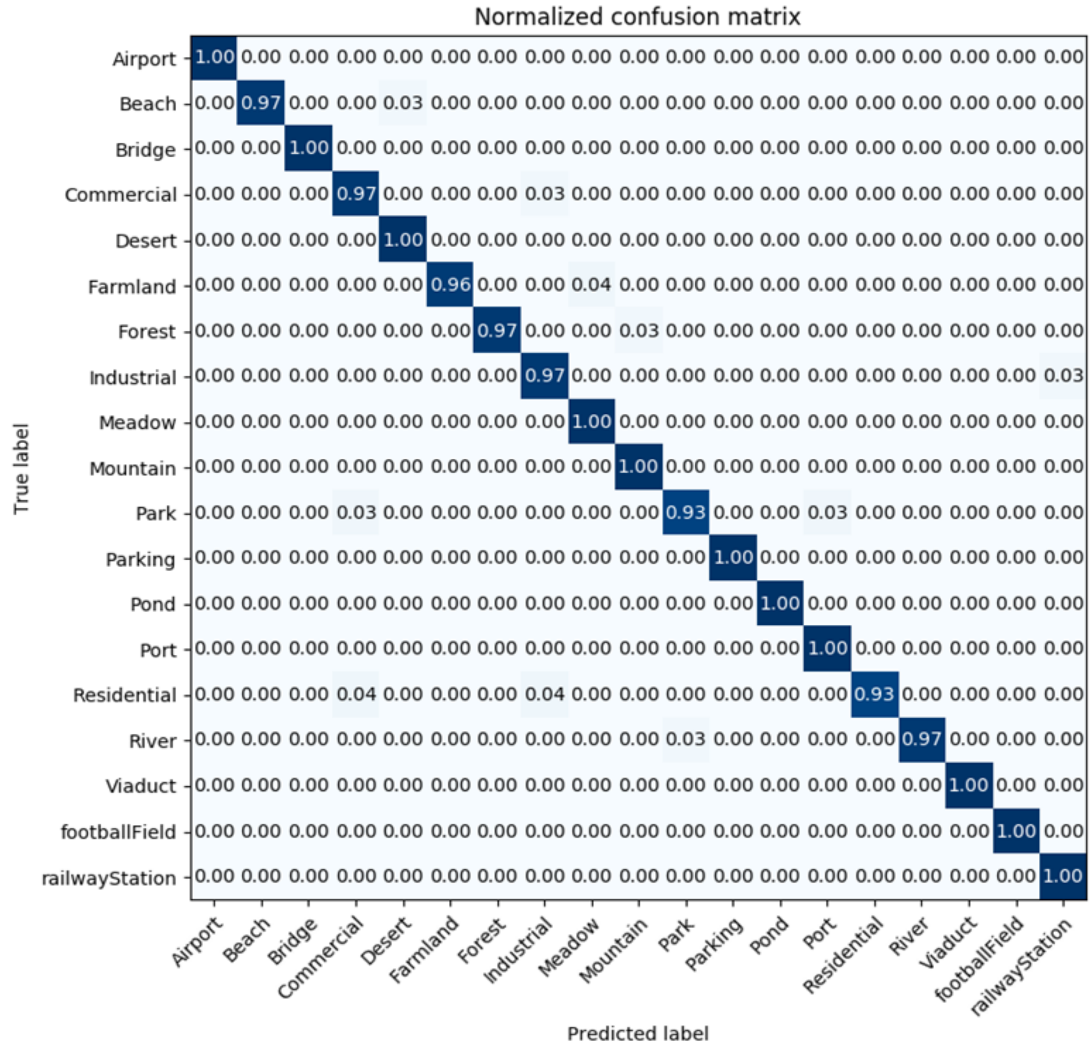
Classification accuracy (OA (%) and SD) of the examined method and the reference methods with 60% and 40% of WHU-RS data set as training set

Results - classification based on features fusion with PCA transformation (10)



Confusion matrix of the examined method with 60% of WHU-RS data set as training set for InceptionV3 mixed_8 (PCA) + ResNet50 avg pooling

Results - classification based on features fusion with PCA transformation (11)



Confusion matrix of the examined method with 40% of WHU-RS data set as training set for DenceNet121 conv5_block16_concat (PCA) + ResNet50 avg pooling

Discussion – feature extraction (1)

- the biggest accuracies are achieved by features extracted from the intermediate convolutional layers
- LRC and SVM gave similar results, challenging task to move further to lower layers
- classification accuracy attained on the class “dense residential” is higher compared to the other classification methods

Discussion – feature extraction (2)

- method of feature fusion with PCA transformation performs better under a smaller percentage of the training set
- data augmentation
- stratified data split may lead to bigger classification accuracies
- Random Forest, XGBoost, Adaboost, or Extremely Randomized Trees

Conclusion

- The proposed technique for remote sensing image classification can be further explored with:

- extracting features from lower layers of pre-trained deep CNN

- stratification of the spilt of training/ testing data set

- Experiments with other small-scale remote sensing data sets, because the proposed classification method gives good results under small training ratio