Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks

Dimitrios Marmanis, Mihai Datcu, *Fellow, IEEE*, Thomas Esch, and Uwe Stilla, *Senior Member, IEEE*

***Abstract*—Deep learning methods such as convolutional neural networks (CNNs) can deliver highly accurate classification results when provided with large enough data sets and respective labels. However, using CNNs along with limited labeled data can be problematic, as this leads to extensive overfitting. In this letter, we propose a novel method by considering a pretrained CNN designed for tackling an entirely different classification problem, namely, the ImageNet challenge, and exploit it to extract an initial set of representations. The derived representations are then transferred into a supervised CNN classifier, along with their class labels, effectively training the system. Through this two-stage framework, we successfully deal with the limited-data problem in an end-to-end processing scheme. Comparative results over the UC Merced Land Use benchmark prove that our method significantly outperforms the previously best stated results, improving the overall accuracy from 83.1% up to 92.4%. Apart from statistical improvements, our method introduces a novel feature fusion algorithm that effectively tackles the large data dimensionality by using a simple and computationally efficient approach.**

***Index Terms*—Convolutional neural networks (CNNs), deep learning (DL), feature extraction, land-use classification, pretrained network, remote sensing (RS).**

I. Introduction

SUPERVISED classification of very high spatial resolution (VHSR) images is still an open research topic in the remote sensing (RS) field. Monitoring urbanization trends has become a crucial objective, and there is currently a high demand for such automatic RS classification techniques. Toward this direction in the last years, advanced methodologies have significantly contributed to the solution of the VHSR classification problem. Predominantly, methods based on the bag-of-visual-words (BoVW) approach have been proposed for solving this task by learning a dictionary for representing the image content in an unsupervised manner through the use of well-established feature descriptors (HOG, SIFT, etc.) and clustering algorithms. Such approaches include the spatial pyramid matching kernel (SPMK) [1], spatial pyramid cooccurrence kernel (SPCK++) [2], min-tree kd-tree [3], and sparse coding [4] methods. However, the main drawback of all these techniques lies in the assumption that a general feature descriptor can adequately represent the complex image structures by employing expert knowledge through manually designed all-around purpose features.

An alternative approach to feature descriptors is unsupervised feature learning through the use of autoencoders and their variants [5]. These methods allow a deep learning (DL) model to learn a set of rich nonlinear representations directly from the input data with no assumptions or prior knowledge. Authors in [6] and [7] have investigated such an approach over RS data; however, their implementations remain shallow due to the use of a single feature extraction level.

The method we adapt suggests using a large pretrained network for general knowledge discovery with no training phase or use of labels. Precisely, the data are forwarded through a large pretrained convolutional neural network (CNN) generating a set of high-level representations, which can be later classified in a second processing stage. Investigations over similar approaches can be found in [8] and [9], where the authors achieved state-of-the-art results over various classification benchmarks. Interestingly, despite the success of such methods in computer vision, there is still no relevant study in the RS domain. For filling this gap in this letter, we propose a novel framework for classifying remotely sensed data using a small set of training labels. Specifically, our approach makes use of a largely trained network over an entirely different classification task, namely, the ImageNet challenge, where raw image data are used as an input in order to generate a set of representations in an unlabeled manner. Then, we forward the derived representations into a CNN classifier along with their respective labels and apply supervised learning. Through this framework, we are able to achieve state-of-the-art results over the UC Merced Land Use (UCML) by improving the overall accuracy from 83.1% to 92.4%. Furthermore, for testing the model's transferability, we perform a set of qualitative experiments over unseen aerial images, by classifying them into our predefined UCML classes and visually assessing the result.

Our contributions in this work are related to the use of large pretrained networks for RS land-use classification, where we show that CNN classifiers from different domains can be profoundly suitable for our classification task. Furthermore, in the framework of DL, we propose a novel technique for fusing representations from multiple hidden layers, which leads to significant dimensionality reduction in addition to an extensive decrease in computation.

Manuscript received June 25, 2015; revised October 14, 2015; accepted October 30, 2015. Date of publication December 1, 2015; date of current version December 24, 2015.

D. Marmanis is with the German Remote Sensing Center (DFD), German Aerospace Center, 82234 Weßling, Germany and also with the Department of Photogrammetry and Remote Sensing, Technical University of Munich, 80333 Munich, Germany (e-mail: Dimitrios.Marmanis@dlr.de).

M. Datcu is with the Remote Sensing Technology Institute (IMF), German Aerospace Center, 82234 Weßling, Germany (e-mail: Mihai.Datcu@dlr.de).

T. Esch is with the German Remote Sensing Center (DFD), German Aerospace Center, 82234 Weßling, Germany (e-mail: Thomas.Esch@dlr.de).

U. Stilla is with the Department of Photogrammetry and Remote Sensing, Technical University of Munich, 80333 Munich, Germany (e-mail: Stilla@tum.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2015.2499239

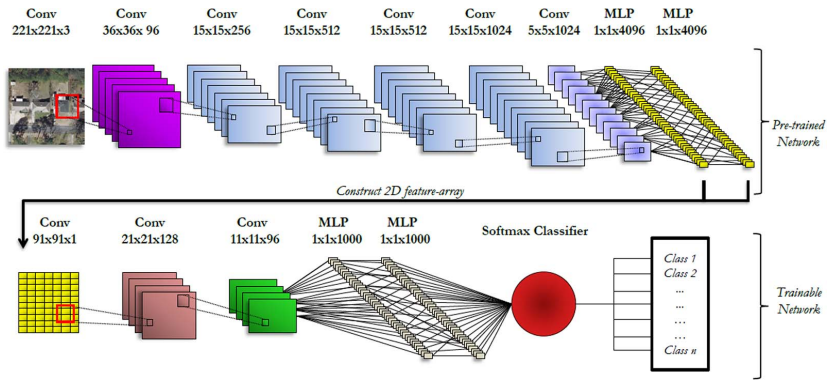

Fig. 1. Workflow diagram of the two-stage classification neural network model.

The remaining parts of this letter are organized as follows. Section II presents an introduction to DL pretrained models. Section III discusses the proposed framework for feature fusion. Section IV introduces our complete pretrained model. Section V presents the experimental framework, whereas Section VI concludes this letter by discussing the results and possible future research directions.

II. Pretrained Models

A. ImageNet Data Set

The ImageNet database is of fundamental importance for vision-related classification tasks as all major pretrained models proposed in the literature are initially trained over it. ImageNet consists of about 15 million high-resolution labeled images separated in roughly 22 000 categories. The images contained in the database are retrieved through search engines and therefore can be considered as common multimedia data. Under the ImageNet framework, a yearly classification competition is held, namely, the "ImageNet Large-Scale Visual Recognition Challenge," where participants make use of a subset of this data set for training classification algorithms of their choice. The actual size of the training data set used in the contest consists of about 1.2 million images, and participants need to classify a test data set into 1000 distinct classes.

B. Supervised Pretrained Networks for RS Image Classification

Using information derived from deep pretrained CNNs on ImageNet, authors in [8] showed that encapsulated representations contained within can work remarkably well for a large set of diverse image classification tasks and often outperform standard classification approaches. These results support the idea that representations from very deep networks are generic and can facilitate transfer learning between different domains, even in cases where limited amount of labeled data are available for the task at hand. Such a property can be much important in fields such as RS, where the availability of labels is scarce and involves extended effort and cost for acquiring (ground-truth campaigns).

An intriguing question is how can RS classification benefit from these systems considering the fundamental differences between the imaging properties of the underlying tasks. In our view, the reasoning for this interconnection lies in the structural arrangement of CNNs, which partially imitates the mammalian visual cortex through a multilayer processing approach. Using such hierarchical architectures, largely trained CNNs accumulate extended knowledge on low-level spatial descriptors (edge and corner detector) and employ them to describe images in bottom-up manner. This way, pretrained CNNs can trivially "decompose" images into a set of primitive elements and detect similarities on various levels of abstraction among them. Due to these incorporated properties, pretrained network can be easily adapted to new visual tasks by minimally readjusting their learned weights.

This hypothesis, if valid, suggests that RS images can directly relate to multimedia data, and therefore, both media can be tackled jointly through identical CNN classifiers. Our observations support this argument, and we present our detailed results in the experimental chapter in this letter.

III. Framework for Feature Fusion

Our major contribution in this work is the formulation of a novel schema for combining features extracted from deep networks using a simple and computationally efficient approach. Our method uses a set of deep activations obtained from different layers of a pretrained CNN and concatenates their information into a new feature vector. This vector is continuously reshaped into a 2-D array (with appropriate zero padding), as shown in Fig. 1. The proposed adaptation is supported by a set of empirical arguments introduced among various DL studies, which we address in the following. Overall, there are three fundamental questions emerging regarding our proposed architectures.

- *Why extract features solely from the very last layers?*
- *Why fuse this information into a single vector?*
- *Why reshape the feature vector into a 2-D array?*

A. Finding Richer Information in Deeper Layers

In the influential work of [8], the authors derived activations from the deep layers of a pretrained CNN and evaluate their

classification accuracy on an individual basis. Through detailed experimentation, they empirically proved that representation contained in the last layers of largely pretrained networks is of major influence toward the classification accuracy, in contrast to the earlier layers, which insignificantly affect the outcome. Therefore, in order to comply with this ascertainment, we uniquely considered information derived from the very deep layers of our pretrained model, completely discarding all shallower features.

B. Feature Fusion for Higher Accuracy

Combining information from different layers of a CNN classifier was initially proposed by [10] and can be perceived as a "multiscale feature extraction," where information of different scales coexists into a single layer within the network. Such an approach is common in DL literature, where skip connections between layers transform a strictly sequential network into a directed acyclic graphs with an overall positive impact on the classification. Another similar approach is shown in [11], where authors concatenated deep features derived from two individual CNNs (trained on RGB and depth images, respectively) into a single vector before passing them to the final classifier. Influenced by these studies, we adapt to the proposed framework by concatenating our deep features into a new single vector and consider them as a joint representation.

C. Spatial Arrangement of High-Level Representation

The intuition behind the reshaping process of the derived feature vector into a 2-D array, depicted in Fig. 1, is motivated by the large dimensionality of our new representation set. Precisely, our current vector has a prohibitively large dimension, which poses a computational bottleneck when considering the optimization of a multidimension objective function in a classification scheme. With this limitation in mind, we reshape our feature vector into a 2-D array, where a large reduction in the number of parameters can occur by employing a second CNN classifier. This reduction emerges due to the fundamental CNN property of sharing weights through a convolution process along its input dimensions. Such a structural adaptation is plausible, as long as sharing parameters among the spatially placed data can produce meaningful associations and detect characteristic patterns among them. In our particular task, each of the elements of the 2-D array can be perceived as a high-level representation that contains general characteristics regarding the class information of each image. Therefore, associations between these high-abstraction image characteristics are credible and should hold valid when considering a shared weights framework. To the best of our knowledge, such a spatial transformation has not been previously proposed in the literature; however, this adaptation seems not to restrict our model to learn and achieve high performance over the UCML classification task.

IV. Model

Our model consists of two individual processing stages, namely, the *pretrained model* and the *trainable CNN model*. In the following, we refer to these classification stages in details, providing information on their properties, characteristics, and architectural design.

TABLE I
Model Architecture and Hyperparameters

Number of kernels per layer	128 / 96
Kernel size - per layer	13x13 / 11x11
Max Pooling size - per layer	3x3 / 3x3
Num. of neurons in FC layers	1000 / 1000
Mini-batch size	10
Learning rates initial / final	0.015 / 0.002
Drop-out - per layer	30 / 50 / 50 / 50%
Weight-Decay	0.00015
Max-Norm	1.9
Momentum initial / final	0.1 / 0.9
Data Augmentation	Zoom, Shear, Rotate, Scale

A. Pretrained Model

The first component of our two-stage classification scheme is the pretrained CNN model, which is employed for generating a set of representations using a fixed set of weights and no labels. For this task, we utilize the Overfeat model [9], which is a popular choice among the publicly available pretrained models. Overfeat is an improved version of the highly reputed AlexNet [12] and is trained on 1.2 million labeled images, which are contained in the 2013 ImageNet data set. Overfeat is provided by its authors as a ready-to-use software, incorporating extended capabilities over deep feature generation and classification.

With respect to implementing the pretrained classification phase, we sequentially feed our UCML training data into Overfeat and extract preactivations from the seventh and eighth hidden layers, respectively. We then concatenate the derived features into 2-D arrays of size 91×91, before passing them into the proceeding trainable classification stage, as shown in Fig. 1. Importantly, all UCML images need to be downsampled from an original size of 259×259 pixels down to 221×221 so that they comply with the standardized input dimension of the Overfeat model, initially determined by its authors.

B. Trainable CNN

The second component of our classification architecture is a trainable CNN that accepts as an input the previously derived 2-D features, along with their respective class labels. The architecture of the trainable CNN contains two convolutional layers, two fully connected layers, and a Softmax classifier on top (see Fig. 1). The model is trained using standard backpropagation and stochastic gradient descent with mini batches, in addition to a set of regularizers, namely, Momentum, Max-Norm, Drop-out, and Weight-Decay, and real-time Data Augmentation. The model hyperparameters are provided in Table I.

The importance of the trainable CNN stage is essential due to its high discriminant potential, the system is able to learn complex relations among highly abstract representations and correctly separate the data into mutually exclusive classes. Despite the fact that other classification algorithms may also deliver accurate results, the CNN significantly outperforms them due to its enhanced spatial adaptation over the input (2-D).

TABLE II
CLASSIFICATION COMPONENTS AND ALGORITHM COMPARISON

Method & Algorithm	Test-set Accuracy
Random Forest with RGB feature	44%
CNN with RGB feature	44.5%
Random Forest with Overfeat features	86.9%
CNN with Overfeat feature	**92.4 %**

TABLE III
METHOD COMPARISON OVER THE UCML BENCHMARK

Method & Algorithm	Test-set Accuracy
BOVW [2]	71.8%
SPMK [1]	74%
SPCK++ [2]	76%
Sparce Coding [4]	81.7%
Salient Unsupervised Learning [6]	82.8% $\pm$ 1.18%
MinTree + KD-Tree [3]	83.1% $\pm$ 1.2%
CNN with Overfeat feature	**92.4 %**

This argument is supported by empirical experiments in the next section.

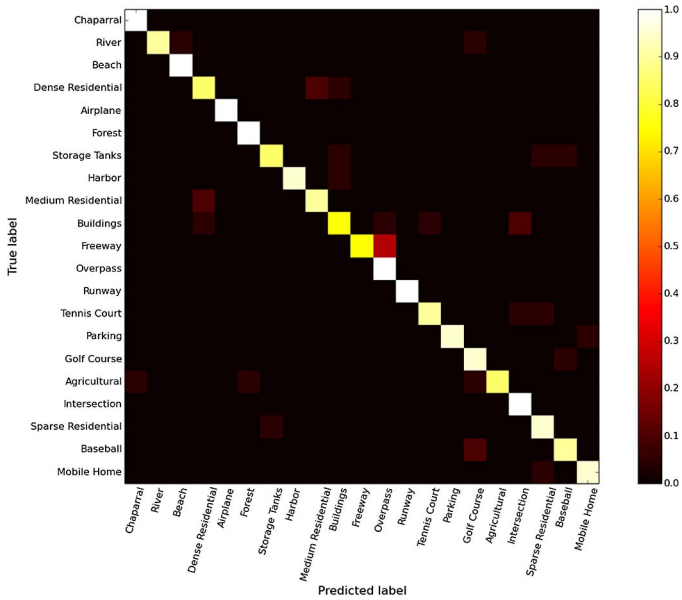

Fig. 2. Confusion matrix of the two-stage classifier over the UCML test data set.

V. EXPERIMENTS

A. Model Evaluation Over UCML Benchmark

UCML is an open-source publicly available aerial image data set of approximately 30-cm spatial resolution, acquired by the U.S. Geological Survey [2]. It contains a total of 2100 image patches of size 256 $\times$ 256 (RGB bands) from various U.S. cities and is separated in 21 semantic categories with 100 instances per class.

By employing our proposed two-stage classifier, we successfully tackle the UCML classification problem, achieving state-of-the-art results with respect to the previously reported outcomes, which are summarized in Table II. For this experiment, we randomly select 70%, 10%, and 20% of the data as training, validation, and test, respectively. We also maintain the class balance to avoid bias in our results. To demonstrate the importance of our two-stage architecture, in Table III, we report classification results when different components of our architecture are omitted or replaced by another well-established classifier, namely, the random forest (RF) algorithm. Our results report that, when the pretrained stage is completely omitted, both the CNN and RF classifiers fail to adequately generalize over the test set, resulting in poor classification performance. Furthermore, through our proposed 2-D spatial arrangement and the use of CNNs along with Overfeat, we significantly improve the classification accuracy compared with the Overfeat-RF architecture, by an absolute difference of 5.5%. The price for this improvement comes from the extended time that the CNN requires for training (approximately three days); however, on test time, both CNN and RF perform equally fast (few seconds). For completeness, it is important to mention that all classification algorithms in Table II achieve a perfect fit of 100% accuracy over the training data set, which indicates the overfit problem that all algorithms exhibit, due to the small size of the training data set. Finally, for evaluating the individual land-use class detection in Fig. 2, we compute the confusion matrix of our two-stage classification model.

B. Investigation of High-Dimensional Feature Structures

An engaging question is how the features are structured into the high-dimensional feature space along the different classification stages. With this in mind, we employed the "t-distributed stochastic neighboring embedding" algorithm (t-SNE) and visualized the derived embeddings into three distinct processing stages, namely, on 1) *RGB level*, 2) *pretrained fused feature level*, and 3) *trained-CNN last layer feature level*. The respective outcomes can be found in Fig. 3. By inspecting the derived clusters, it is clear that the prior information obtained by the pretrained network positively affects the data and leads to an initial separation in mutually exclusive classes, as shown in Fig. 3(b). Similarly, through the employment of the second trainable network, further disentanglement occurs by increasing the separability and relative distance between the individual semantic clusters.

C. Model Transferability Over New Data

One essential milestone for RS remains the model transferability for classifying unseen data derived from different sensors and geographical areas. For tackling this problem, we employ our elaborated UCML model, over a new set of manually extracted aerial images of same resolution from various cities (Boston, USA; San Diego, USA; Valladolid, Spain). Despite the complexity of the unseen data analyzed, the model appears to deliver accurate results as it correctly detects the classes depicted in the images in Fig. 4. These outcomes support the argument that our model has constructed a set of high-level representations that can accurately detect the underlying semantic objects, even in significantly diverse scenes.

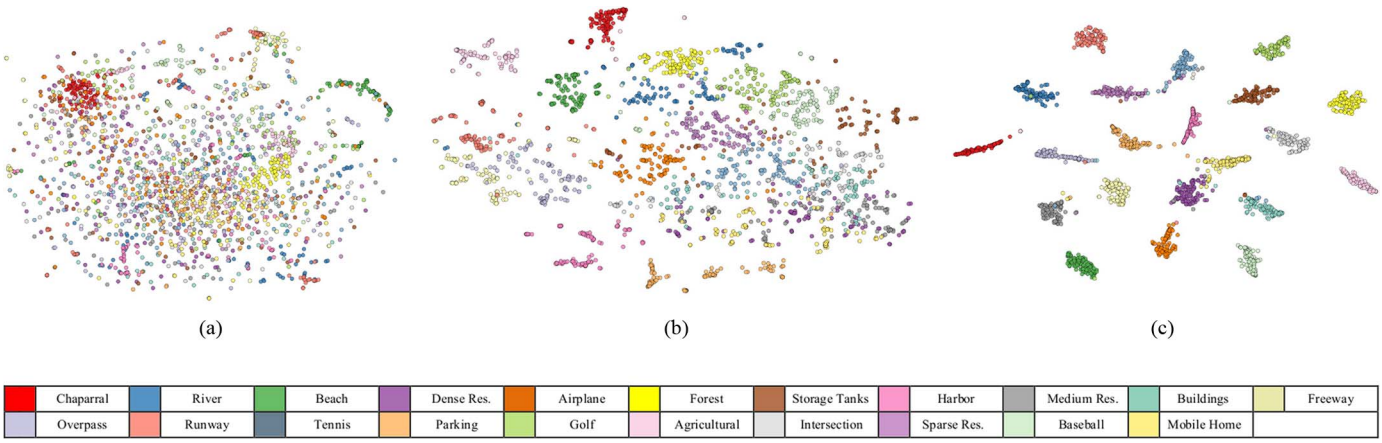

Fig. 3. Two-dimensional scatterplots of high-dimensional features generated with t-SNE over the UCML data. (a) Scatterplot of RGB pixels as features. (b) Scatterplot of features extracted from last two layers of ImageNet-Overfit network. (c) Features extracted from last supervised CNN. All points in the scatterplots are class coded.

Fig. 4. Predicted classes over unseen aerial data.

VI. Discussion and Conclusion

In this letter, we have investigated the potential of using large pretrained neural networks, for classifying RS aerial images into a large set of diverse land-use classes. Through our proposed framework, we have achieved promising results over the UCML benchmark, significantly increasing the best stated performance through a simple and computationally efficient end-to-end approach.

In our future research, we are planning to investigate the potential of pretrained networks on a larger scale experiment, considering satellite data with greater spectral resolution and geographical variations.

References

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2169–2178.

[2] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE ICCV*, 2011, pp. 1465–1472.

[3] L. Gueguen, "Classifying compound structures in satellite images: A compressed representation for fast queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1803–1818, Apr. 2015.

[4] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[6] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[7] O. Firat, G. Can, and F. T. Yarman Vural, "Representation learning for contextual object and region detection in remote sensing," in *Proc. 22nd IEEE ICPR*, 2014, pp. 3708–3713.

[8] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," unpublished paper, 2013. [Online]. Available: http://arxiv.org/abs/1310.1531.

[9] P. Sermanet *et al.*, "Overfeat: Integrated recognition, localization and detection using convolutional networks," unpublished paper, 2013. [Online]. Available: http://arxiv.org/abs/1312.6229

[10] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Proc. IJCNN*, 2011, pp. 2809–2813.

[11] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 665–673.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.