# Very High Resolution Images Classification by Fusing Deep Convolutional Neural Networks

**Meziane Iftene[a,*], Qingjie Liu[b] and Yunhong Wang[c]**

State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, 37 Xueyuan Road, Haidian District, Beijing 100191, China
[a] m_iftene@hotmail.fr, [b] qingjie.liu@buaa.edu.cn, [c] yhwang@buaa.edu.cn

**Keywords:** CNN Fusion, Convolutional neural networks, Fine-tuning, VHR images, Classification.

**Abstract.** Recently, deep convolutional neural networks (CNNs) have made great achievements, whether taken as features extractor or classifier, in particular for very high resolution (VHR) images classification task which is a key point in the remote sensing field. This work aims to improve the VHR image classification accuracy by exploiting the fusion of two pre-trained deep convolutional neural network models. In this paper, we propose to concatenate the features extracted from the last convolutional layer of each pre-trained deep convolutional neural network to get a long features vector which is fed into a fully-connected layer and then perform a fine-tuning for a VHR image dataset classification. The experimental results are promising since they show that the fusion of two deep CNNs achieves better accuracy for the classification compared to the individual CNN models on the same dataset.

## 1. Introduction

The technological advancements have made deep convolutional neural networks [1] outperform other algorithms by an order of magnitude in classification tasks, especially for very high resolution remotely sensed image field and this, despite the large number of complex land cover classes which makes its efficient classification a challenging problem. The deep convolutional neural network contains a large number of parameters, and therefore they are trained with very large datasets. Despite recent advancements in capturing remote sensing imagery by satellite, drones and planes that made the availability of very high resolution images, this field still lacks of large datasets to correctly train a deep CNN.

To overcome this problem, one approach is to use the CNN as features extractor and Razavian et al. [2] demonstrated that features extracted from the pre-trained CNN should be considered as a generic image representation in most visual recognition tasks. For example, Hu et. al [3] investigated transferring the activations of pre-trained CNNs to high resolution images classification task by extracting features from the last convolutional layer and fully-connected layer of a deep network pre-trained on ImageNet as generic features and further, coded them using traditional feature encoding methods like BOW [4], locality-constrained linear coding (LLC) [5],

Vector of Locally Aggregated Descriptors (VLAD) [6] and Improved Fisher Kernel (IFK) [7] to generate a robust image representation and they achieved remarkable accuracies. Another approach proposed by Oquab et. al [8], is to fine-tune the CNN parameters using the target dataset after that parameters are learned on a large dataset with data diversity such as ImageNet database [9]; this approach is called Transfer Learning (TL) and the parameters transfer success depends on the similarity of the target dataset to the ImageNet database. Taking this into account in the previous work [10], we evaluate the generalization of deep convolutional neural networks features from the last fully-connected layer in the classification process of a small dataset of a VHR imagery by fine-tuning CNN deeper layers only. The obtained results show better accuracy compared to that of CNNs as features extractor.

Focused on continued improvement of CNN efficiency and in order to overcome the lack of a large dataset, a number of recent techniques in recognition task have proposed to fuse two features extractors, and the first notable progress was made by Simonyan et al [11] who proposed a two-stream network architecture designed to mimic the pathways of the human visual cortex for object and motion recognition. Very recently, bilinear CNNs were proposed [12], using two CNNs pre-trained on the ImageNet dataset and fine-tuned on fine-grained image classification datasets. The outputs of two CNNs are multiplied using the outer product at each location of the image, and are pooled to obtain an image descriptor.

In this paper, we propose fusing two pre-trained CNN models for VHR image classification where, the outputs from the last convolutional layer of each network are concatenated to get a feature vector, then we perform a fine-tuning for the classification by adding fully connected layers on top as shown in figure 1. The experiment is performed on WHU-RS dataset [11]. The experimental result shows that the fusion of two CNNs obtains better classification accuracy compared to the individual CNNs fine-tuned or taken as features extractors on the WHU-RS dataset.

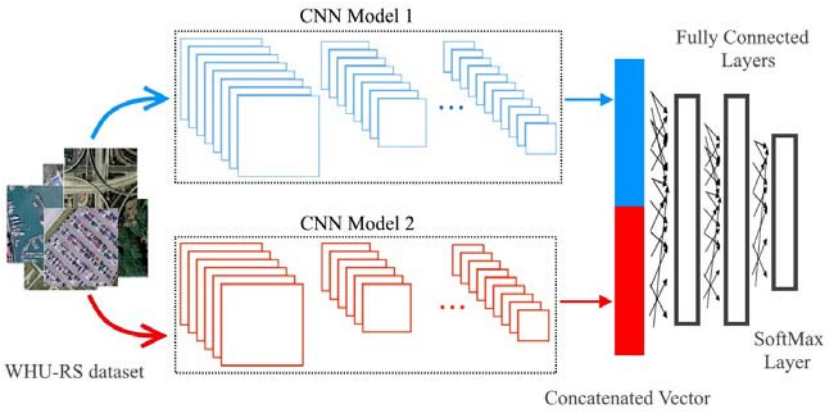


Figure 1 Illustration of two CNN models fusion.

## 2. Proposed Approach

The fusion illustration showed above is inspired by the bilinear models proposed by Tenenbaum and Freeman [14] to model the separation between "content" and "style" factors of perceptual systems, and is motivated by the promising results obtained using the bilinear CNNs applied to fine-grained categorization [12].

This work is based on two popular CNNs, CaffeNet [15] and GoogleNet [16] whose parameters were trained on 1,200,000 ImageNet images and 1,000 object classes as features extractors. The first model is a replication of the AlexNet model [17], it contains 5 convolutional layers and 3 fully

connected layers with a minor difference in the order of the pooling and the normalization layers (in CaffeNet, pooling is done before normalization). The second model is deeper, it contains an Inception module which enables Google's team to go as deep as 22 layers, but at the same time has even far less parameters (12 times less) than AlexNet model.


Figure 2 Examples of each category in the WHU-RS dataset [12].

Both networks are considered as features extractors where, features are extracted from the last convolutional layer of CaffeNet and the output of the top inception module of GoogleNet then, the features are concatenated to create a long vector which is fed into fully-connected layers followed by a k-way SoftMax layer initialized randomly where, k is the number of classes of the target dataset. Finally, we perform a fine-tuning of the model for the classification of WHU-RS dataset, which contains 50 satellite images with a size of 600×600 for each of the 19 classes, collected from Google Earth. The collection from different resolution satellite images, the variation in scale, illumination and viewpoint-dependent appearance in some categories make this dataset more complicated [3] as displayed in Figure 2. The results obtained are evaluated by comparing them to those obtained by [10], where authors fine-tuned the CNN models, and [3] where the CNNs are considered as features extractors; both works were carried out for the classification of the WHU-RS dataset.

## 3. Experiments and Results

In this paper, we evaluate our approach with different combination of CaffeNet and GoogleNet: the first one is initialized with two CaffeNet models, the second is initialized with a CaffeNet and a GoogleNet models and the third one is initialized with two GoogleNet models. The features are extracted using both networks, then features are concatenated and the entire model is fine-tuned for several epochs at decreased learning rate while the SoftMax activation function is trained at a normal rate, preserving the parameters of the previous models and transferring that learning into our new model. The experiments have been performed on the deep learning framework Caffe developed by Yangqing Jia [18]. Following the form of 5-fold cross validation, we perform training and testing split.

The results obtained for the three combinations of CaffeNet and GoogleNet as features extractors from the last convolutional layer (C5) and the output of the top inception module (IM) respectively, are reported for WHU-RS dataset in the following tables where they are compared to the results

obtained for the same dataset classification task by fine tuning each CNN model [10] and CaffeNet model as features extractor from fully connected layer and convolutional layer [3].

Table 1 Comparison with previously obtained accuracies on the WHU-RS dataset with GoogleNet.

| | Fine-tuning [10] | | Fusion with IM | Fusion with C5 |
|---|---|---|---|---|
| | Without data augmentation | With data augmentation | | |
| GoogleNet | 94.72 % | 96.32% | **97,65** | **98,22%** |

Table 2 Comparison with previously obtained accuracies on the WHU-RS dataset with CaffeNet.

| | Fine-tuning [10] with data augmentation | Feature extractor [3] | | | Fusion with IM | Fusion with C5 |
|---|---|---|---|---|---|---|
| | | 1st FC layer | 2nd FC layer | Features coded with IFK method | | |
| CaffeNet | 96.84% | 96.23% | 95.58% | 97.43% | **98,22%** | 97,31% |

Tables 1 and 2 display the WHU-RS dataset classification accuracies of our different fusion models compared with the previous work done on WHU-RS dataset. “IM” and “C5” refer to the concatenation with features extracted from the output of the top inception module of GoogleNet and from the last convolutional layer of CaffeNet respectively. “FC” refers to Fully Connected layer. Thus, “Features coded with IFK method” refers to features extracted from the last convolutional layer of CaffeNet and coded with Improved Fisher Kernel method. The bolded values are the highest of these classifications results.

The results presented above show that our fusion of CaffeNet and GoogleNet without any pre-processing or data augmentation for the WHU-RS dataset classification task improves the accuracy from 2% to 3.5% compared with individual GoogleNet without and with data augmentation. Even for the fusion of two GoogleNet we obtained a bit higher accuracy (1%) compared to GoogleNet fine-tuned with data augmentation. The fusion of two CaffeNet improves the results obtained for the compared methods and got almost the same accuracy as the features coded with IFK method. Compared to CaffeNet fine-tuned with data augmentation, the fusion of CaffeNet and GoogleNet improves the accuracy with 1.5%, 2% compared with CaffeNet as features extractor from 1st FC layer, 3% compared with CaffeNet as features extractor from 2nd FC layer and almost 1% compared with the features of the convolutional layer coded with IFK.

## 4. Conclusions

In this novel work, we propose the fusion of two pre-trained deep convolutional neural networks to improve the classification accuracy of the WHU-RS dataset. The results illustrate that the fusion of two CaffeNet models (C5+C5) turned out to be the lowest performing fusion model with 97.31% accuracy while, fusing two GoogleNet models (IM+IM) performed slightly better accuracy with 97.65%. The best result is achieved by fusing two different models CaffeNet and GoogleNet (C5+IM) with an accuracy of 98.22%. The results show that the fusion of two different models outperformed the individual models; this can be explained by the fact that each network architecture leads to an image representation which can differ slightly from an architecture to another even if each network is trained on the same dataset. Even the lowest performing fusion model with same two networks (C5+C5 with 97.31%) performed almost the same accuracy as the best performing individual model taken as features extractor and codded with IFK method (97.43%).

## References


[1] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” Handb. brain theory neural networks, 3361(10), (1995).

[2] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: an astounding baseline for recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 806–813 (2014).

[3] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, “Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery,” Remote Sens., 7(11), 14680–14707 (2015).

[4] J. Sivic, A. Zisserman, and others, “Video google: A text retrieval approach to object matching in videos.,” in iccv, 2(1470), 1470–1477 (2003).

[5] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in Computer Vision and Pattern Recognition (CVPR), IEEE Conference, 3360–3367 (2010).

[6] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” IEEE Trans. Pattern Anal. Mach. Intell., 34(9), 1704–1716 (2012).

[7] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in European conference on computer vision, 143–156 (2010).

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1717–1724 (2014).

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, Proc. CVPR. IEEE Conference on, 248–255 (2009).

[10] M. Iftene, Q. Liu, and Y. Wang, “Very high resolution images classification by fine tuning deep convolutional neural networks,” in Eighth International Conference on Digital Image Processing (ICDIP), 100332D-100332D (2016).

[11] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems, 568–576 (2014).

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE International Conference on Computer Vision, 1449–1457 (2015).

[13] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, “Structural high-resolution satellite image indexing,” in ISPRS TC VII Symposium-100 Years ISPRS, 38, 298–303 (2010).

[14] J. B. Tenenbaum and W. T. Freeman, “Separating style and content with bilinear models,” Neural Comput., 12(6), 1247–1283 (2000).

[15] Y. Jia et al., “Caffe: Convolutional architecture for fast feature embedding,” in Proc. ACM International Conference on Multimedia, 675–678 (2014).

[16] C. Szegedy et al., “Going deeper with convolutions,” arXiv Prepr. arXiv, 1409-4842 (2014).

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 1097–1105 (2012).

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proc. ACM International Conference on Multimedia, 675–678 (2014).