International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2017, Submitted in January 2017, iajesm2014@gmail.com

Study On the Transfer Learning in Aerial Scene Classification

Sri Sharanabasappa Raikoti, Assistant Professor, Department of Computer Science, Government Degree College Yadgir, Karnataka, India, Email: sr.raikoti@gmail.com

Abstract

Remote Sensing (RS) imageclassification has recently attracted great attention for its application in different tasks, including environmental monitoring, battlefield surveillance, and geospatial object detection. The best practices for these tasks often involve transfer learning from pre-trained Convolutional NeuralNetworks (CNNs).

AcommonapproachintheliteratureisemployingCNNsforfeatureextraction, and subsequently train classifiers exploiting such features. In this paper, we propose the adoption of transfer learning by fine-tuning pre-trained CNNs for end-to-end aerial image classification. Our approach performs feature extraction from the fine-tuned neural networks and remote sensing image classification with a Support Vector Machine (SVM) model with linear and Radial Basis Function (RBF) kernels. To tune the learning rate hyperparameter, we employ a linear decay learning rate scheduler as well as cyclical learning rates. Moreover, in order to mitigate the overfitting problem of pre-trained models, we apply label smoothing regularization. For the fine-tuning and feature extraction process, we adopt the Inception-v3 and Xception inception-based CNNs, as well the residual-based networks ResNet50 and DenseNet121. We present extensive experiments on two real-world remote sensing image datasets: AID and NWPU-RESISC45. The results show that the proposed method exhibits classification accuracy of up to 98%, outperforming other state-of-the-art methods.

Keywords: Learning Technique, Remote Sensing, Aerial Scene

INTRODUCTION

The categorization of scene images into a distinct set of meaningful groups based on the image contents is important in the analysis of imaging sensor, aerial and satellite images because of its importance in an extensive range of applications, significant various approaches for remote sensing data scene classification have been developed throughout the last few decades. The capacity to handle large dimensionality data and perform well with limited training samples, as well as high accuracy, drew a lot of attention with the introduction of SVM machine learning. We can use a pretrained network as a commencement for learning a new deep learning task by replacing the pretrained network's fully connected layer and classification layer.

Transfer learning helps in easing the training process as the trainable network parameters got reduced which can avoid overfitting. This method is mostly adopted on small datasets due to inability to train huge parameters of deep neural network architecture. Several feature selection or feature extraction methods have been developed so far based on transferred features from pre-trained deep neural network which are already trained on large image datasets like Imagenet. Imagenet is a natural image dataset where images are captured horizontally unlike remote-sensing images which are captured from the sky. Though there is structural differences between both the type of images but still the models that are pre-trained on imagenet can be transferred for remote-sensing scene classification task. The reason behind this is the local similarity between natural images and remote sensing images is higher as proved in, hence the convolutional filters of earlier layers of pre-trained networks can more precisely describe local structure information in remote sensing images whereas the descriptions of thedeeper convolutional layers are more meaningful.

In general, features of the last fully connected layer of a CNN are taken forclassification purpose but the features extracted from mid layers of CNN or any convolutional layer may also have some significance in classification. Moreover, convolutional layer features are more discriminative than fully connected layer features as the former contain more spatial information than the later. Hence, in the work, adaptive deep pyramid matching (ADPM) model is proposed that combines the features from all of the convolutional layers by a convolution fusion technique. In paper, the novel multi-model feature extraction networkcombines multiple pretrained CNN models to extract the features

of images. In TEX-Nets, pre-trained deep learning CNN architectures are encoded with Local Binary Patterns (LBP) which is a handcrafted texture descriptor for texture recognition to develop texture coded mapped images. These texture coded mapped images are used to train TEX-Nets which provide complementary information to the standard RGB deep models by fusing it with the standard RGB stream. The work considers the last convolutional layer of a pre-trained CNN models as multi-scale feature extractor where the features after extraction are encoded using sparse coding to achieve scene classification. Again, pre-trained deep CNNs are used as feature extractor to extract deep features of aerial images from different network layers and then fed into the Support Vector Machine (SVM) for classification. Similarly, the paper explores the benefits of multi-layer features by extracting features from multiple layers of pre-trained CNN and integrates them using a fusion strategy called PCA/SRKDA for improving the scene classification in different aspects. Two-stage deep feature fusion model also combines features from different layers to generate two converted CNNs which are based on two well-known CNN architectures and then fused them to further improve the classification performance.

Main Approaches

The two main approaches to adapt pre-trained networks as found in literature are:

(a) Pre-trained networks are used as feature extractors. After applying training data on pre-trained CNN, features are extracted from a desired layer totrain a classifier.

(b) Fine-tuning the whole pre-trained network or some of the layers using target data and then features are extracted to train the classifier. This strategy is basically applied in deeper layers of pre-trained deep CNNs to further improve the classification performance by freezing lower layers and allowing higher layers to learn by training them with target data.

The reason behind fine-tuning of the higher layers of pre-trained CNN is that the low level features can better fit remote sensing images as these features are more generic when compared to high-level features. High-level features are specific to a particular dataset hence the fine-tuning with that particular target dataset helps in improving classification results. In paper two pre-trained CNN architectures namely, CaffeNet and GoogleNet are adopted using both the above approaches for aerial image classification. For effective training process, GoogLeNet employs auxillary classifiers in the intermediate layers of the network and also applies filters of various sizes in each layer to get more accurate spatial information of an image. Again, this work exploits pre-trained CNN models to extract an initial set of representations and then transferred into a supervised CNN classifier to avoid to extensive overfitting due to availability of limited training data. Pre-trained Inception v5 architecture is used to generate feature vectors of remote sensing images in work which are then applied to train a random-forest-based classifier. Similarly, features extracted from pre-trained networks are used to train non-neural network classifiers like SVM, KNN classifier and random forest in paper [10] and in second approach, a softmax classifier is added at the end of the pre-trained network and fine-tune all the parameters by retraining with target dataset. In work CNN-CELM method is proposed where CNN is used as feature extractor and all the deep convolutional feature vectors are normalized before fed to CELM-based classifier for land-use scene classification. The semi-supervised deep rule-based (SSDRB) approach also employs pre-trained CNN for extraction of high level features from the sub-regions of the images. After an efficient supervised initialization process using few labeled training images, meta-parameters are self-updated from the unlabeled images in a fully unsupervised manner. With the aim of reducing the overfitting, a novel framework based on the Siamese convolutional neural networks with rotation invariance regularization has been developed in work.

Apart from focusing on feature extraction using pre-trained deep CNNs, some methods have been developed based on other metrics like preprocessing of input data, type of classifier used, way of encoding the extracted features and many more for successful transferring pre-trained models to classification task. For successful transfer learning task, a linear

PCA network (LPCANet) is de- signed in work [77] to synthesize spatial information of remote sensing images in each spectral channel before transfer learning process and quaternion algebra toLPCANet is introduced to synthesize spectral information. A multi-scale feature extraction approach based on pre-trained CNN models is proposed in paper [34] where features from the last convolutional layer with respect to different scales ofthe images are encoded using BOW and Fisher encoding to create global feature representations for aerial images. The global features of the images are fed into a classifier for the scene classification task. Unlike previous CNN-based methods which ignores local objects of images, an end-to-end CNN model is proposed to capture both Global-Context features (GCFs) and Local-Object level fres (LOFs). The concatenation of both GCFs and LOFs produces a feature set which is more discriminative for classification than only GCFs. Objects in remote-sensing images are relatively small compared to objects in natural images so it is hard to detect both GCFs and LOFs using existing pre-trained CNNs and this issue is solved in this paper. Fully connected layers at the end of CNN does not capture hierarchical features in images which is important for classification. Hence, the last convolutional layer of a pretrained deep CNN model is selected as feature extractor in to create initial features and then these features are fedinto CapsNet that works on spatial information of features in an image to generatefinal feature set for classification. The hybrid Deep CNN (DCNN) feature classifi- cation by ensemble extreme learning model (EELM) is proposed where three parallel pre-trained heterogeneous DCNN models are used for feature extraction to generate hybrid features using linear connection of each of three feature set. To improve the discriminative power of the feature sets, joint loss function is applied while training these parallel models individually. Again in another method, two pretrained convolutional neural networks (CNNs) are used as feature extractor, the first one extract features from original aerial image and the second from the processed aerial image. Each of the feature set are then fused through two feature fusion strategies. In a work, two approaches are adopted for aerial classification: firstly, off-the-shelf pre-trained CNN model is used to extract high dimensional features of aerial images followed by a traditional classifier and the other approach is to retrain a pre-trained CNN model using aerial images thatis known as finetuning, and applied the fine-tuned network directly for classifica- tion on target images. Scene classification using deep neural networks is a very time-consuming process particularly during training. Hence, CNN architecture, RSSCNet which is a no-freezing transfer learning method has been proposed to speed up the training process with improved classification accuracy.

Proposed method

In this paper, we evaluate four different CNN architectures to solve the problem of high-resolution aerial scene classification. We adopt CNNs that are pre-trained on the ImageNet dataset with the purpose of determining their effectiveness in remote sensing image classification tasks. First, we explore the fine-tuning of the weights on the aerial image dataset. In the process of fine-tuning, we remove the final layers of each of the pre-trained networks after the average pooling layer (so called “network surgery”) and construct a new network head. The new network head consists of: a fully connected layer, dropout, and a softmax layer. Network training is performed on the modified deep neural network. Subsequently, we exploit fine-tuned CNNs for feature extraction and utilize the extracted features for the training of SVM classifiers, which have been successfully applied in other image classification and transfer learning problems. In this paper, SVMs are implemented in two versions: with linear kernel and with Radial Basis Function (RBF) kernel. We use a linear decay learning rate schedule and cyclical learning rates and evaluate their suitability for fine-tuning of pre-trained CNNs for remote sensing image classification. Moreover, we apply label smoothing as a regularization technique and assess its impact on the classification accuracy compared with state-of-the-art methods. Figure 1 shows a flowchart of the proposed method. The main contributions of this paper are (1) evaluation of modern CNNs models on two remote sensing image datasets, (2) analysis of the impact of linear learning rate decay

schedule and cyclical learning rates from the aspect of classification accuracy, (3) evaluation of label smoothing on model generalization compared to state-of-the-art techniques, and (4) assessment of the transferability of the features obtained from fine-tuned CNNs and their classification with linear and RBF SVMs classifiers. To the best of our knowledge, the combination of adaptive learning rate and label smoothing was never studied before in the context of aerial scene classification.

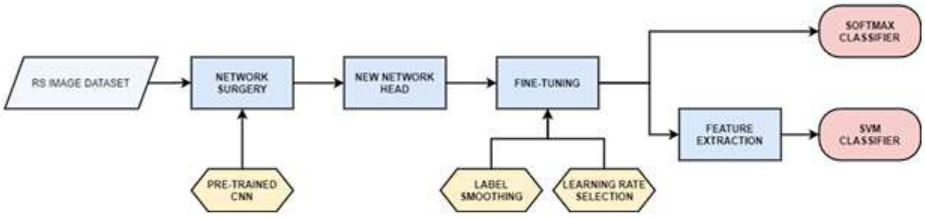

Figure 1. Flowchart of the proposed method.

Conclusion

CNN-based classification is the most popular state-of-the-art classification ap- proach in aerial scene classification. Since aerial scenes are RGB images thus they can be easily fit into a CNN due to its' 3D input dimension. In any classification task, the performance depends on the quality of extracted feature set. Several fac- tors can negatively affect the discriminative power of the feature set in aerial scene classification such as poor learning of network parameters due to either network overfitting or vanishing gradient, small inter class variations and large intra class variations, the inefficient way of extracting features and many more. Throughout the literature it has been found that transfer learning performs better than any other deep learning based frameworks for scene classification as it helps in better parameter optimization by avoiding overfitting on small datasets like aerial scenes. Apart from this, the concept of Few-shot learning has come out that also works well on small datasets. To improve the discriminative power of the feature set, attention mechanism has been incorporated in CNNs in several works such that more attention can be given to some special areas of an image that can give more discriminative features rather than the entire image. The aim of this thesis is to improve classification performance on aerial scene datasets which is fulfilled by tackling two research issues network overfitting and vanishing gradient problem. Hence, all the contributory works here are focusing on either of the two issues.

Reference:

Bazi, Y.; Rahhal, M.M.A.; Alhichri, H.; Alajlan, N. Simple Yet Effective Fine-Tuning of Deep CNNs Using an Auxiliary Classification Loss for Remote Sensing Scene Classification. Remote Sens. 2016, 11, 2908.

Bengio, Y. *et al.* Greedy layer-wise training of deep networks. *Advances in neural information processing systems* **19**, 153, 2007.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recog-nition*. 1251–1258, 2016.

Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 2015, 7, 1468014707.

Lee, C.-Y. *et al.* Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Artificial intelligence and statistics*. 464–472, PMLR, 2016.

Lu, Y.; Luo, L.; Huang, D.; Wang, Y.; Chen, L. Knowledge Transfer in Vision Recognition. ACM Comput. Surv. 2016, 53, 1–35.

Penatti, O.A.; Nogueira, K.; dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; p. 4451.

Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Columbus, OH, USA, 23–28 June 2014; p. 512519.

Song, Z. *et al.* A sparsity-based stochastic pooling mechanism for deep convolutional neural networks. *Neural Networks* **105**, 340–345, 2016.

Tong, Z. *et al.* A hybrid pooling method for convolutional neural networks. In *International Conference on Neural Information Processing*. 454–461, Springer, 2016.

Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.