Transfer Learning in Aerial Scene Classification

Dr. Prateek Mishra, Professor, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Rabia Shaheen, Research Scholar, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Email- rabiashaheenghulam@gmail.com

Abstract: 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.

Keywords: Learning Technique, Remote Sensing, Aerial Scene

INTRODUCTION

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 [38], 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.

The two main approaches to adapt pre-trained networks as found in literature [69] 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 to train 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 [70]. 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 [69], 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 [71] 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 [71] 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 classifierand 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 [11], 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 [72] 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 [73].

In general, features of the last fully connected layer of a CNN are taken for classification 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 [37], 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 [74], the novel multi-model feature extraction network combines multiple pretrained CNN models to extract the features of images. In TEX-Nets [39], 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 [39]. The work [12] 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 [40] 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 [75] 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 [76] 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.

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 to LPCANet 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 of the 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 [78] is proposed to capture both Global-Context features (GCFs) and Local-Object level features (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 [78]. Fully connected layers at the end of CNN does not capture hierarchical features in images which is important for classification [79]. Hence, the last convolutional layer of a pretrained deep CNN model is selected as a feature extractor in [79] to create initial

features and then these features are fed into CapsNet that works on spatial information of features in an image to generate final feature set for classification. The hybrid Deep CNN (DCNN) feature classifi- cation by ensemble extreme learning model (EELM) [80] 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 [80]. Again in another method [17], 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 [10], 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 that is 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 [81] which is a no-freezing transfer learning method has been proposed to speed up the training process with improved classification accuracy.

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:

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, 2017.

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.

Song, Z. *et al.* A sparsity-based stochastic pooling mechanism for deep convolutional neural networks. *Neural Networks* **105**, 340–345, 2018.

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