Review of Literature on The Deep Learning Techniques for Classifying Remote Sensing Aerial Scenes

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

Earth observation is a collection of information about the earths surface whether it be physical, chemical and biological systems using earth observation satellite or earth remote sensing satellite or directly captured from aircrafts. With the increasing volume of high-resolution remote-sensing images due to development of such earth observation technologies, there necessitate automated systems for analyzing as well as classification of these images for various applications like land mapping, vegetation and so on. Earlier image classification is done based on hand- crafted features that require human intervention which is quite difficult to handle huge number of data. With the advent of deep learning, researchers took the opportunity of incorporating deep neural networks into image classification model to ease the feature extraction process. The trend of deep learning is growing up day by day as the researchers fully focused on developing new deep learning tech- niques and trying to outperform existing ones.

Keywords: Review of Literature, Learning Technique, Remote Sensing

INTRODUCTION

Spatial resolution is a measurement of how clearly visible or detailed objects are in an image based on pixels. Pixel is the smallest unit of an image which are combined to form an image. That means, an image of size 32 × 32 × 3 has totalof 2352 number of pixels. A spatial resolution of 200 m means that one pixel represents an area 200 by 200 meters on the ground. Therefore, images with high spatial resolution have smaller pixel size and that of lower spatial resolution have larger pixel size. High resolution images have more detailed objects with more number of pixels compared to low resolution images. Cameras are used as sensors in aerial photography and photos are taken from the sky with the help of helicopters, aircraft, and spacecraft. The ground coverage of an aerial photo depends on several factors, including the focal length of the lens, the platform altitude, and the format and size of the film. The focal length effectively controls the angular field of view of the lens and determines the area covered by the camera. The longer the focal length, the smaller the area covered on the ground, but with greater detail. The area covered also depends on the altitude of the platform. Higher the altitude, larger is the area covered by the camera on the ground but with reduced detail and lower the altitude, the smaller is the area covered on the ground, but with greater detail. That means, when the altitude is high, then the images captured by cameras are lower resolution images which can have no positive impact on good classification performance. Some researchers have empirically studied these factors that can impact the accuracy of aerial image classification. Most of the studies have focused on the impact of changes in spatial resolution in very high-resolution aerial images for classification task. This study [8] compared the impact of spatial resolution on land/water classifications and found that a small-magnitude change (11.5 m) in spatial resolution has negligible impact on the classification performance. Again this study [9] evaluated the impact of satellite image spatial resolution on land use classification and found the classification accuracy was 82.3% for spatial resolution of 1 m and 75.1% for spatial resolution of 30 m.

Due to availability of high resolution aerial scenes, this research work purely focused on feature selection as well as classification frameworks for en- hancing the classification performance.

Deep learning and its architecture

Deep learning is a branch of machine learning established by Hinton and Salakhut- dinov [10][11][12][13] in 2006. Hierarchical features of the input data (e.g. Image) are computed, where the higher-level features are obtained by combining the lower- level ones. Deep learning use the architecture of a deep neural network shown in figure 2-1, which is a multilayer network with several hidden layers of nodes be- tween input and output, whose weights are initialized after training process. The layers between input and output do feature identification and processing in a series of stages by increasing the level of abstraction from one layer to another.

Review of Literature

Despite of getting impressive results in deep learning methods, the accuracies on public aerial scene datasets have almost reached saturation due to less availability of training samples. Therefore, to improve the scene classification performance, continuous efforts have been given to promote new methods in scene classification task. Aerial scene classification using transfer learning have been gaining fame gradually, where intermediate features extracted from pre-trained model are employed for image representation in classification task. A lot of methods are developed for feature extraction using pre-trained deep CNNs. For example, different layers of a CNN extract different level of features of an image. But in most works, the features from the last fully-connected layers are taken for classification task ignoring the other convolutional layer features which may also help in getting good classification results. Several works [37][38][39][12][40] have been proposed where either convolutional features orfeatures from fully-connected layers are employed for remote sensing image classi- fication. Apart from transfer learning methods, there exist several deep learning based frameworks developed by researchers each having the aim of outperforming earlier published methods in terms of classification accuracy. Generally, while extracting features of an image, the entire image area is considered in most ofthe methods ignoring the fact that the only discriminative regions are essential for extracting powerful discriminative features for scene classification. Hence, the concept of Attention Mechanism has been evolved to give more importance to such regions of an image as well as feature maps of CNNs and implement in several works [41][42][43][44][45][46][47][48][49][50][51][52][53][54][55][56][57][56].

In a work of scene classification [41], a new architecture referred to as Unified Attention Network (UAN) was proposed that learns to attend to different Convolutional Neural Networks (CNN) layers and specific layers within a given feature map in a sequential manner that combines the "what level of abstraction to attend to, and "where should the network look at different parts of the inputs.Two steps are done: Feature selection using Soft Attention and Layer selection using Hard Attention. At each LSTM timestep, the soft attention mechanism selects a feature that can improve task performance by probing through the input image to effectively classify multiple objects. The hard attention mechanism selects the layer whose output achieves best task performance. That means, output of the selected CNN layer is processed by spatial soft attention atdifferent LSTM steps. Again, in another work [42], attention image is generated using Grad-CAM architecture and the high-feature from original images and attention map are fused using an inconsistent two-stream architecture joint optimization. A global-local attention network (GLANet)[52] is proposed to capture both global and local information for aerial scene classification unlike existing CNN based models that neglect the local information about scenes. Both global information and the local semantic information are learned through attention mechanisms. To discriminate the key components and their semantic relationship inside a scene, this paper [53] uses attention mechanism in the proposed novel remote sensing scene classification method based on high-order graph convolutional network (H-GCN). In H-GCN, the information about the neighbor nodes are computed at different orders which made each node more informative. High-order self-attention network (HoSA) is proposed in a work [54] where a self-attention module captures long-range dependencies within the scenes for extracting high-level semantic features. After that, high-order pooling mechanism is applied to further explore high-order information present in the features. Similarly, Attention Recurrent Convolutional Network (ARCNet) [55] to adaptively select some key regions or locations where the recurrent attention structure does the extraction of high-level semantic and spatial features. To reflect the importance of complex objects in remote sensing scenes as well as focusing more on the infrequently occurring features, self-attention-based deep feature fusion (SAFF) [56] has been developed which aggregates multi-layer features extracted from a pretrained convolutional neural network (CNN) model with the help of spatial-wise weighting and channel-wise weighting. Spatial-Wise Weighting focuses on the characteristics of complex objects of the scenes and channel-wise weighting focuses on the differences among the images by increasing the weights of infrequently occurring features. An end-to-end model, CAE-CNN[57] also uses attention mechanism to capture the most discriminative feature by focusing the most class-specific region in each aerial scenes. Attention mechanism based on CNN focus on discriminative regions of an image, but it may suffer from the influence of intra-class diversity for which an attention-based deep feature fusion (ADFF) [56] framework is proposed.

CONCLUSION

Designing neural network architecture also plays an important role in suc-cessful feature extraction as well as classification task. It is not necessary that a deep architecture works best in any type of datasets. For example, in handwritten digits recognition, a swallow neural network is more suitable than the deeper ones. In a paper [58], a CNN based architecture known as ERNet is developed for aerial scene classification of disaster events with less depth and low-computational that is suitable for low-cost devices. The running time of this architecture is up to 3 times faster on an embedded platform and provides similar accuracy to existing models with less than 2% accuracy drop compared to the state-of-the-art. Several well-known CNN architectures are developed so far which require a considerable amount of architecture engineering skills and domain knowledge. All these archi- tectures may not give their maximum benefits in all type of remote sensing scene datasets. Hence, in a paper [59], an automatic architecture learning procedure is proposed for remote sensing scene classification where the optimal set of parame- ters (every set of parameters such as convolution, pooling etc. represent different CNN architecture) for a given dataset have been learned by means of gradient descent. To extract more abstract semantic information from aerial images, the network must be deeper which also increases parameters. Hence, in a work [60], full convolutional network based on DenseNet is designed that constructs a small number of convolutional kernels to generate a large number of reusable feature maps by dense connections that can increase the network depth without increase the number of parameters significantly. In a paper [61], three different convo- lutional neural networks with different sizes of receptive field are designed and fused with a proposed probability fusion model which is multilevel fusion method that gives the label of the image. In work [62], two novel modules such as skip connections and covariance pooling are embedded into the traditional convolu- tional neural network (CNN) model. Here, feature maps of three convolutional layers are combined using skip connections to obtain multi-resolution feature maps and covariance pooling aggregates the multi-resolution feature maps that helps to achieve more representative feature learning in remote sensing datasets. Similar concept is applied in another work [63], where multilayer feature maps, obtained by pretrained convolutional neural network (CNN) models are stacked together and a covariance matrix is calculated for the stacked features to find the covari- ance of two different feature maps. Finally, these covariance matrices represent the features for classification task. Generally, CNNs are directly applied on an entire scene image without considering the scale variation of the objects in the aerial scenes that can be a factor of wrongly classified scene images. This is solved in a work [64] where patches of an image with random size are taken and stretched to a specific size as input to the CNN taken for classification purpose. This deep random-scale stretched CNN extracts features that are robust to the scale varia- tion. Another possible factor is small inter class variations and large intra class variations in aerial scenes. Dilated Convolutional Neural Network (D-CNN) [65] tries to overcome this issue that improves the performance of aerial scene classification by increasing the receptive field of convolutional layer without increasing parameters. The need of large amounts of training data in deep CNN creates dif- ficulty among small aerial scene dataset due to improper feature learning. Apart from transfer learning, Few-shot learning [66] can learn a model using only a few labeled samples. A meta-learning method is proposed for few-shot classification of aerial scene images where the framework consists of a feature extractor, a meta- training stage, and a meta-testing stage. The feature extractor is trained on allbase categories to learn a representation of inputs and then in the meta-training stage, a meta-learning classifier is optimized in the metric space by cosine dis- tance with a learnable scale parameter [66]. Similarly to address this issue, a novel convolutional neural network named JM-Net [67] is designed that has fewer parameters since different size of convolution kernels are applied in same layer unlike fully convolutional layer. Deep color models [68] of CNN are developed by exploring different color spaces and their combination and all of them are fused to investigate the importance of color within the deep learning framework for aerial scene classification. This proves its' effectiveness by improving the classification performance compared to using only the RGB image as input to the network as a general practice. Two novel deep architectures, texture coded two-stream deep architecture and saliency coded two-stream deep architecture which are based on the idea of feature-level fusion are proposed in a work [57] to further improve the classification accuracy.

Reference:

Anwer, R. M. *et al.* Compact deep color features for remote sensing scene classification. *Neural Processing Letters* **53** (2), 1523–1544, 2021.

Castelluccio, M. *et al.* Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092* , 2015.

Lu, X. *et al.* Jm-net and cluster-svm for aerial scene classification. In *IJCAI.* 2386–2392, 2017.

Luo, C. *et al.* Utilization of deep convolutional neural networks for remote sensing scenes classification. In *Advanced Remote Sensing Technology for Synthetic Aperture Radar Applications, Tsunami Disasters, and Infrastruc- ture,* IntechOpen, 2018.

Marmanis, D. *et al.* Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters* **13** (1), 105–109, 2015.