The Deep Learning Techniques for Classifying Remote Sensing Aerial Scenes

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

Abstract

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 to further improve the classification accuracy. Remote sensing image scene classification with deep learning (DL) is a rapidly growing field that has gained significant attention in the past few years. While previous review papers in this domain have been confined to 2015, an up-to-date review to show the progression of research extending into the present phase is lacking. In this review, we explore the recent articles, providing a thorough classification of approaches into three main categories: Convolutional Neural Network (CNN)-based, Vision Transformer (ViT)-based, and Generative Adversarial Network (GAN)-based architectures. Notably, within the CNN-based category, we further refine the classification based on specific methodologies and techniques employed. In addition, a novel and rigorous meta-analysis is performed to synthesize and analyze the findings from 50 peer-reviewed journal articles to provide valuable insights in this domain, surpassing the scope of existing review articles. Our meta-analysis shows that the most adopted remote sensing scene datasets are AID (41 articles) and NWPU-RESISC45 (40). A notable paradigm shift is seen towards the use of transformer-based models (6) starting from 2015. Furthermore, we critically discuss the findings from the review and meta-analysis, identifying challenges and future opportunities for improvement in this domain. Our up-to-date study serves as an invaluable resource for researchers seeking to contribute to this growing area of research.

Keywords: remote sensing; **deep learning**; **scene classification**; **convolutional neural networks**; **meta-analysis**

INTRODUCTION: 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. 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 total of 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 in 2006. 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, 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 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 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) 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) 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) framework is proposed.

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 multi-layer 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. 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 have been proposed where either convolutional features or features 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 of the 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.

Understanding Remote Sensing Image Scene and Deep Learning Feature Extraction

High-Resolution Scene Classification Datasets

Multiple VHR remote sensing datasets are available for scene classification. The UC Merced Land Use Dataset (UC-Merced or UCM) [**51**] is a popular dataset obtained from the United States Geological Survey (USGS) National Map Urban Area Imagery collection covering US regions. Some of the classes belonging to this dataset are airplanes, buildings, forests, rivers, agriculture, beach, etc. The Aerial Image Dataset (AID) is another dataset acquired from Google Earth Imagery with a higher number of images and classes than the UCM dataset. Some of the scene classes are common in multiple datasets. For instance, the “forest” class is included in UCM, WHU-RS19, RSSCN7, and NWPU-RESISC45 datasets. However, there are variations in scene classes among different datasets. In addition, the number of images and their size and resolution of scene datasets varies in respective datasets. Thus, the selection of the dataset depends on the research objectives. Cheng et al. proposed a novel large-scale dataset NWPU-RESISC45 with rich image variations and high within-class diversity and between-class similarity, addressing the problem of small-scale datasets, lack of variations, and diversity. Miao et al. merged UCM, NWPU-RESISC45, and AID datasets to prepare a larger remote-sensing scene dataset for semi-supervised scene classification and attained a similar performance to the state-of-the-art methods. In these VHR datasets, multiple DL architectures have been conducted to obtain optimal accuracy.

Pretrained CNNs for Feature Extraction: Collecting and annotating data for larger remote sensing scene datasets increases costs and is a laborious task. To address the scarcity of data in the remote sensing domain, researchers often utilize terrestrial image datasets such as ImageNet and PlacesCNN [**67**], which contain a large number of diverse images from various categories. Wang et al. described the local similarity between remote sensing scene image and natural image scenes. By leveraging pretrained models trained on these datasets, the CNN algorithms can benefit from the learned features and generalize well to remote sensing tasks with limited labeled data.

In 2015, Penatti et al. [Discovering better classification results for the UCM dataset than low-level descriptors. In a diverse large-scale dataset named NWPU-RESISC45, three popular pretrained CNNs: AlexNet, VGG-16 and GoogLeNet, improved the performance by 30% minimum compared to handcrafted and unsupervised feature learning methods. NWPU-RESISC45 dataset is ambiguous due to high intra-class diversity and inter-class similarity. Sen et al. adopted a hierarchical approach to mitigate the misclassification. Their method is divided into two levels: (i) all 45 classes are rearranged into 5 main classes (Transportation, Water Areas, Buildings, Constructed Lands, Natural Lands), and (ii) the 45 sub-levels are trained in each class. DenseNet-121 pretrained on ImageNet is used as a feature extractor in both levels. Al et al. combined four scene datasets, namely UCM, AID, NPWU, and PatternNet, to construct a heterogeneous scene dataset. For suitability, the 12 shared classes are filtered to utilize an MB-Net architecture, which is based on pretrained ResNet-50. MB-Net is designed to capture collective knowledge from three labeled source datasets and perform scene classification on a remaining unlabeled target dataset. Shawky et al. brought a data augmentation strategy for CNN-MLP architecture with Xception as a feature extractor. Sun et al. obtained multi-scale ground objects using multi-level convolutional pyramid semantic fusion (MCPSF) architecture and differentiated intricate scenes consisting of diverse ground objects.

Yu et al. introduced a feature fusion strategy in which CNNs are utilized to extract features from both the original image and a processed image obtained through saliency detection. The extracted features from these two sources are then fused together to produce more discriminative features. Ye et al. proposed parallel multi-stage (PMS) architecture based on the GoogleNet backbone to learn features individually from three hierarchical levels: low-, middle-, and high-level, prior to fusion. Dong et al. integrated a Deep Convolutional Neural Network (DCNN) with Broad Learning System (BLS) for the first time in the remote sensing scene classification domain to extract shallow features and named it FDPResNet. The DCNN implemented ResNet-101 pretrained on ImageNet as a backbone on both shallow and deep features and further fused and passed to the BLS system for classification. CNN architectures for remote sensing scene classification vary in design with the incorporation of additional techniques and methodologies. However, the widely employed approaches include LBP-based, fine-tuning, parameter reduction, and attention mechanism methods.

LBP-based pretrained CNNs: LBP is a widely used robust low-level descriptor for recognizing textures. In 2015, Anwer et al. proposed Tex-Net architecture, which combined an original RGB image with a texture-coded mapped LBP image. The late fusion strategy (Tex-Net-LF) performed better than early fusion (Tex-Net-EF). Later, Yu et al., who previously introduced the two-stream deep fusion framework, adopted the same concept to integrate the LBP-coded image as a replacement for the processed image obtained through saliency detection. However, they conducted a combination of previously proposed and new experiments using the LBP-coded mapped image and fused the features together. Huang et al. [**84**] stated that two-stream architectures solely focus on RGB image stream and overlook texture-containing images. Therefore, CTFCNN architecture based on pretrained CaffeNet [**85**] extracted three kinds of features: (i) convolutional features from multiple layers, wherein each layer improved bag-of-visual words (iBoVW) method represented discriminating information, (ii) FC features, and (iii) LBP-based FC features. Compared to traditional BoVW [**35**], the iBoVW coding method achieved rational representation.

Fine-tuned pretrained CNNs: Cheng et al. not only used pretrained CNNs for feature extraction from the NWPU-RESISC45 dataset, they further fine-tuned the increasing learning rate in the last layer to gain better classification results. For the same dataset, Yang et al. fine-tuned parameters utilized on three CNN models: VGG-16 and DenseNet-161 pretrained on ImageNet used as deep-learning classifier training, and feature pyramid network (FPN) pretrained on Microsoft Coco (Common Objects in Context) for deep-learning detector training. The combination of DenseNet+FPN exhibited exceptional performance. Zhang et al. used the hit and trial technique to set the hyperparameters to achieve better accuracy. Petrovska et al. implemented linear learning rate decay, which decreases the learning rate over time, and cyclical learning rates. The improved accuracy utilizing fine-tuning on pretrained CNNs validates the statement made by Castelluccio et al. in 2015.

Parameters reduction: CNN architectures exhibit a substantial amount of parameters, such as VGG-16, which comprises approximately 138 million parameters. The large number of parameters is one of the factors for over-fitting. Zhang et al. utilized DenseNet, which is known for its parameter efficiency, with around 7 million parameters. Yu et al. integrated light-weighted CNN MobileNet-v2 with feature fusion bilinear model and termed the architecture as BiMobileNet. BiMobileNet featured a parameter count of 0.86 million, which is six, eleven, and eighty-five times lower than the parameter numbers reported in and , respectively, while achieving better accuracy. from entire images, it is essential to consider that images contain various objects and features. Therefore, selectively focusing on critical parts and disregarding irrelevant ones becomes crucial. Zhao et al. added a channel-spatial attention module (CBAM) following each residual dense block (RDB) based on DenseNet-101 backbone pretrained on ImageNet. CBAM helps to learn meaningful features in both channel and spatial dimensions . Ji et al. proposed an attention network based on the VGG-VD16 network that localizes discriminative areas in three different scales. The multiscale images are fed to sub-network CNN architectures and further fused for classification. Zhang et al. introduced a multiscale attention network (MSA-Network), where the backbone is ResNet. After each residual block, the multiscale module is integrated to extract multiscale features. The channel and position attention (CPA) module is added after the last multiscale module to extract discriminative regions. Shen et al. incorporated two models, namely ResNet-50 and DenseNet-121, to fulfill the insufficiency of single CNN models. Both models captured low, middle, and high-level features and combined them with a grouping-attention-fusion strategy. Guo et al. proposed a multi-view feature learning network (MVFL) divided into three branches: (i) channel-spatial attention to localize discriminative areas, (ii) triplet metric branch, and (iii) center metric branch to increase interclass distance and decrease intraclass distance. Zhao et al. designed an enhanced attention module (EAM) to enhance the ability to understand more discriminative features. In EAM, two depthwise dilated convolution branches are utilized, each branch having different dilated rates. Dilated convolutions enhance the receptive fields without escalating the parameter count. They effectively capture multiscale contextual information and improve the network's capacity to learn features. The two branches are merged with depthwise convolutions to decrease the dimensionality. Hu et al. introduced a multilevel inheritance network (MINet), where FPN based on ResNet-50 is adopted to acquire multilayer features. Subsequently, an attention mechanism is employed to augment the expressive capacity of features at each level. For the fusion of features, the feature weights across different levels are computed by leveraging the SENet approach.

Reference:

Anwer, R. M. *et al.* Compact deep color features for remote sensing scene classification. *Neural Processing Letters* **53** (2), 1523–1544, 2015.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.

Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-normalizing neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *30*, 971–980.

Lin, D.; Fu, K.; Wang, Y.; Xu, G.; Sun, X. MARTA GANs: Unsupervised representation learning for remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *14*, 2092–2096.

Lu, X. *et al.* Jm-net and cluster-svm for aerial scene classification. In *IJCAI*. 2386–2392, 2015.

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, 2015.

Ma, D.; Tang, P.; Zhao, L. SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro. *IEEE Geosci. Remote Sens. Lett.* **2014**, *16*, 1046–1050.

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

Neupane, B.; Horanont, T.; Aryal, J. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sens.* **2011**, *13*, 808

Xu, S.; Mu, X.; Chai, D.; Zhang, X. Remote sensing image scene classification based on generative adversarial networks. *Remote Sens. Lett.* **2014**, *9*, 617–626.