Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/101/20240932

Remote sensing image scene recognition using MobileNet

Yadong Sun[1], Wenjie Zhao[1,3,*], Jiawen Liu[2]

[1]Navigation College, Dalian Maritime University, Dalian, China
[2]School of Maritime Economics and Management, Dalian Maritime University, Dalian, China

[3]23330402@qq.com
[*]corresponding author

Abstract. Remote sensing image scene recognition plays a pivotal role in various applications, including environmental monitoring, disaster response, urban planning, precision agriculture, and aids in resource management and policy formulation. However, utilizing established convolutional neural networks(CNNs) models like AlexNet and VGG9 for this task can be computationally intensive and time-consuming due to their extensive parameter requirements. This dissertation introduces a MobileNet-based CNN optimized for remote sensing image scene recognition. This lightweight model significantly reduces computational load and model size without compromising accuracy, thereby enhancing efficiency. Empirical results on the NWPU45 dataset demonstrate MobileNet's superiority, achieving an accuracy of 91.16%, a Kappa coefficient of 90.96%, and an F1 score of 91.16% on the test set. Moreover, MobileNet's compact architecture, with merely 3.2531 million parameters and 587.9342 million FLOPs, underscores its efficiency and makes it a promising candidate for practical deployment in remote sensing applications. The findings suggest that MobileNet not only addresses the challenge of computational intensity but also opens new avenues for advancing scene recognition technology in the field of remote sensing.

Keywords: Convolutional neural networks, Deep learning, Scene recognition.

1. Introduction

Remote sensing image scene recognition holds significant practical importance and broad application prospects across various sectors[1]. As a critical earth observation tool, it offers substantial social and economic benefits by promptly detecting environmental changes, enhancing the efficiency of disaster response, and playing a pivotal role in urban planning, land management, and providing vital insights for military strategy.

Traditional neural network models, such as the radial basis function (RBF) network[2] and the feed-forward neural network[3], face challenges in remote sensing image scene recognition. They require manual feature engineering, struggle to capture the complex and high-order relationships within remote sensing imagery, and involve considerable computational overhead.

High-resolution remote sensing imagery, while rich in spatial and textural information, presents a daunting task for scene recognition due to its high complexity, varied imaging conditions, and limitations in spectral resolution. This complexity makes the direct extraction of scene information

DOI: 10.54254/2755-2721/101/20240932

from vast datasets particularly challenging, thus highlighting the difficulty and importance of the task, which has garnered considerable attention in the field.

Current research often focuses on enhancing the accuracy of remote sensing image scene recognition, yet it frequently overlooks the computational demands. This oversight can impede real-time processing and the rapid delivery of recognition outcomes, which are essential for practical applications.

Advances in neural network applications for remote sensing image scene recognition have demonstrated the effectiveness of Recurrent Neural Networks (RNNs)[4], Deep Belief Networks[5], and convolutional neural networks(CNNs)[6]. CNNs, in particular, stand out due to their robust feature extraction capabilities, adaptability, generalization, and adept handling of large datasets, which are advantageous for remote sensing image recognition tasks. However, two primary issues remain: the extensive parameter computation and the time-consuming nature of complex CNN structures, which hinder real-time recognition capabilities.

This paper introduces MobileNet[7], a lightweight CNN that employs Depthwise Separable Convolution(DSC) to alleviate the computational and parameter burden associated with traditional CNNs. Despite its efficiency, MobileNet maintains competitive accuracy levels compared to other mainstream models.

Our contributions are as follows:

1. We have applied CNNs to the task of remote sensing image scene recognition, significantly enhancing the accuracy and efficiency of the recognition process.

2. We have adopted the MobileNetV1 architecture, a lightweight CNN model that effectively extracts features from remote sensing images using DSC, significantly reducing the model's computational demands.

3. To substantiate the effectiveness of our approach, we conducted comparative experiments between MobileNetV1 and other prevalent CNN models on the NWPU45[8] dataset. The results indicate that MobileNetV1 outperforms other models on this dataset, upholding classification accuracy.

The remainder of this paper is structured as follows: Section 2 details the MobileNet model, Section 3 presents the experimental findings, Section 4 provides a summary of our conclusions,and section 5 presents the cited references.

2. Method

2.1. Convolutional neural network

CNNs are a class of deep learning models that have found extensive application in the domains of image recognition and analysis. They draw inspiration from the human visual system, utilizing a series of trainable convolutional kernels to perform convolutional operations that adeptly extract features from images. Following the convolutional layers, pooling layers are strategically employed to reduce the dimensionality of the features and to enhance their invariance to variations, which in turn helps to minimize computational complexity and mitigate the risk of overfitting. Subsequently, a fully connected layer consolidates these processed features to carry out sophisticated classification or regression tasks, culminating in the model's final output.

In the context of remote sensing image scene recognition, the task is fraught with challenges that include managing the vast volume of data, discerning complex surface features, accounting for variable imaging conditions, dealing with a scarcity of labeled data, and tackling the issue of category imbalance. CNNs rise to these challenges by offering an automatic feature extraction mechanism that requires no manual intervention. Their inherent strong translation invariance ensures robustness against shifts in the visual data. The end-to-end learning process of CNNs streamlines the workflow by seamlessly integrating feature extraction and decision-making stages. Furthermore, their excellent generalization capability enables them to perform reliably across diverse and unseen scenarios, making CNNs a powerful tool for remote sensing image scene recognition.

DOI: 10.54254/2755-2721/101/20240932

2.2. Framework

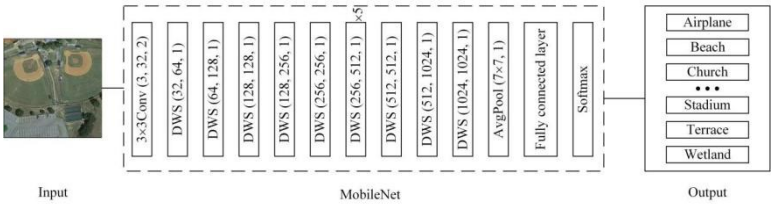

Figure 1. Framework of Scene recognition method.

In the realm of remote sensing image scene recognition, this study leverages the lightweight CNN, MobileNet, and applies it to the NWPU45 dataset—a collection comprising 45 distinct scene categories. This dataset presents several challenges, including high category diversity, sample imbalance, and inconsistent image resolutions across its contents. MobileNet's architecture is adept at efficiently extracting pivotal features from remote sensing images, thereby facilitating successful scene classification with its low computational overhead. Utilizing this approach, we have systematically assigned each image in the dataset to one of the 45 corresponding scene category labels with high precision. This not only underscores MobileNet's effectiveness in handling complex remote sensing imagery but also demonstrates its utility in practical applications where efficiency and accuracy are paramount.

2.3. Depthwise separable convolution

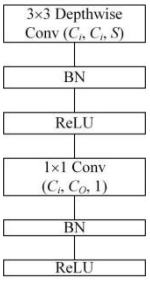

Figure 2. Depthwise separable convolution. DSC(*Ci, Co, S*).

DSC is an efficient optimization strategy designed to mitigate the computational intensity and parameter volume associated with CNNs. This approach intelligently bifurcates the standard convolution process into two more manageable stages: depthwise convolution and pointwise convolution.

Initially, each input channel engages in an independent depthwise convolution through a dedicated 3×3 kernel, which operates without cross-channel interaction. The result is an output channel set identical in number to the input channels (C_i), with spatial dimensions condensed according to a specified step size (S). Subsequent to the depthwise convolution, batch normalization(BN) is applied to expedite the training process and bolster the model's robustness, succeeded by a ReLU activation function that enriches the model's capacity for capturing nonlinear relationships.

Subsequently, pointwise convolution is executed, where a 1×1 convolution kernel is employed to amalgamate the outputs from the depthwise convolutions. This synthesis results in a unified set of output channels (C_0), without any alteration to the spatial dimensions of the feature map. The pointwise convolution allows for channel count modulation, enabling the network to discern composite patterns across various channels. Following this step, BN is reapplied, complemented by another instance of the ReLU activation function.

This ingenious design enables DSC to markedly curtail the parameter count and computational demands of the CNN, without substantially compromising on accuracy. Consequently, it renders the

DOI: 10.54254/2755-2721/101/20240932

model exceptionally well-suited for environments constrained by limited computational resources, offering a judicious balance between efficiency and performance.

3. Experiment

3.1. Dataset

The NWPU45[8]dataset, an esteemed repository for remote sensing image scene classification, originates from Northwestern Polytechnical University and is crafted for pixel-level classification endeavors. It encompasses a diverse spectrum of 45 distinct surface scenes, spanning from airports and baseball stadiums to deserts and farmlands, encapsulating both urban and natural terrains. The dataset offers a wealth of imagery, with each category featuring 700 high-quality images to enrich the research corpus.

The dataset is characterized by its extensive category diversity, uneven sample distribution across classes, and a variance in image resolutions. To ensure consistency, all images have been standardized to a uniform resolution of 256 × 256 pixels. The Ground Sampling Distances (GSDs) exhibit a broad spectrum, ranging from as fine as 0.2 meters to as coarse as 30 meters. In terms of dataset distribution, 80% of the images are allocated for training purposes, with the remaining 20% reserved for testing. This partitioning is strategically designed to facilitate the development and assessment of deep learning models. A visual representation of the NWPU45 dataset can be found , which provides a schematic overview of its composition and structure.

3.2. Experimental setup

In this research, we harnessed the power of NVIDIA RTX3090 GPUs and selected Python 3.10.0 as our development environment, complemented by version 2.3 of the PyTorch framework for the training of our deep learning models. During the preliminary phase of experimentation, we meticulously tuned a suite of hyperparameters to enhance model performance.

Specifically, we designated a total of 60 epochs for the model's training cycle, ensuring ample exposure to the training data. Concurrently, we set the learning rate at 0.0002, a choice that allows us to meticulously control the magnitude of model weight updates, thereby striking a balance between convergence velocity and the fidelity of optimization.

For each iteration of gradient descent, we opted for a Batch size of 16 samples, leveraging the GPU's parallel processing prowess while also optimizing memory utilization. Furthermore, we employed the cross-entropy loss function as our model's loss criterion. This selection is deliberate, as the cross-entropy function adeptly directs the model's optimization trajectory during training. It does so by juxtaposing the probability distributions predicted by the model against the actual labels of the dataset, with the ultimate goal of minimizing the potential for classification errors. This strategic choice of hyperparameters and functions has been instrumental in refining our model's predictive capabilities within the scope of this study.

3.3. Evaluation metrics

In this study, we have employed a comprehensive set of evaluation metrics[9]to appraise the performance of our model across various dimensions. The metrics selected for the classification aspect include Accuracy, which is the ratio of correctly predicted instances to the overall sample count; the Kappa coefficient, a measure that quantifies classification accuracy while accounting for chance agreement; and the F-1 Score, which harmoniously combines the precision and recall of the model to provide a nuanced assessment of its performance.

These metrics are derived from the testing phase, where we calculate the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The observed concordance rate, denoted as p_0, and the expected concordance rate, denoted as p_e, are also determined. TP represents the count of samples accurately identified as belonging to the positive class, while TN signifies the count of samples correctly identified as belonging to the negative class. FP refers to the instances

DOI: 10.54254/2755-2721/101/20240932

where the model incorrectly classifies samples as positive, and FN denotes the instances where the model incorrectly classifies samples as negative.

The observed concordance rate,p_0, is the proportion of samples that are correctly classified by the model relative to the total number of samples. Conversely, the expected concordance rate, p_e, is the proportion of samples that would be expected to be correctly classified if the classification were to occur at random. These metrics collectively offer a robust framework for evaluating the efficacy and reliability of our model within the context of remote sensing image scene recognition.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Kappa = \frac{p_0-p_e}{1-p_e} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{5}$$

3.4. Comparison results

Table 1. Results of popular models.

Models	Accuracy (%)	Kappa (%)	F1 -score (%)	Params (M)	FLOPs (M)
LeNet	59.70	58.78	59.53	7.1613	71.4566
AlexNet	64.67	63.86	64.88	76.0549	1.4978
VGG9	85.08	84.74	85.13	558.5261	14,274,406.4
ResNet18	88.64	88.38	88.56	11.3195	1.8237
MobileNet	**91.16**	**90.96**	**91.16**	**3.2531**	**587.9342**

To substantiate the efficacy of MobileNet in the domain of remote sensing image scene recognition, we conducted a comparative analysis with several other prevalent CNNs. The synthesized experimental outcomes, as delineated in Table 1, indicate that MobileNet outperforms all other models under consideration. The sequence of performance starts with MobileNet, followed by ResNet18[10], VGG9[11], AlexNet[12], and LeNet[13].

ResNet18 also exhibits commendable performance, which can be attributed to its more profound network architecture. This deeper structure is instrumental in mitigating the degradation of features. Despite its capability to furnish a robust feature representation, ResNet18 is encumbered by a substantial count of parameters, leading to escalated demands on computational resources.

In stark contrast, MobileNet presents a dual advantage of a significantly reduced parameter count and high experimental accuracy, thereby achieving a lightweight and efficient feature extraction process. The deeply separable convolutional design of MobileNet enables a substantial reduction in both the parameter count and computational requirements of the model. This design does not compromise the model's ability to adeptly capture the spatial features of images. MobileNet's proficiency in efficient feature extraction, despite its minimalistic parameter profile, is a testament to its effectiveness in remote sensing image scene recognition tasks.

In sum, MobileNet manifests substantial superiority in the sphere of remote sensing image scene recognition, attributable to its trifecta of being lightweight, efficient, and accurate. These attributes make MobileNet a highly promising candidate for deployment in scenarios where computational resources are at a premium, without sacrificing the quality of recognition performance.

DOI: 10.54254/2755-2721/101/20240932

4. Conclusions

This paper introduces a lightweight CNN model, leveraging the MobileNet architecture, for the task of scene recognition in remote sensing imagery. When pitted against other extant CNN models in experiments utilizing the NWPU45 dataset, MobileNet not only exemplifies superior classification accuracy but also excels in Kappa coefficient and F1-score metrics, concurrently achieving a substantial reduction in both parameter count and computational demand. These findings validate that MobileNet is adept at diminishing the computational and storage burdens of the model, without compromising on the precision of recognition. Consequently, the integration of MobileNet into the domain of remote sensing image scene recognition adeptly addresses the challenge of the extensive parameter size inherent in traditional models, while concurrently enhancing computational efficiency, offering an innovative and effective solution for this application.

Looking ahead, we are committed to broadening the application spectrum of the MobileNet model, particularly within the realm of cloud image recognition. Through structural refinements and meticulous parameter tuning, we aim to elevate the model's recognition capabilities. Given the high dimensionality and spatio-temporal intricacies of cloud imagery, we intend to delve into novel strategies for feature extraction and context fusion. Our objective is to bolster the precision of cloud classification, thereby equipping the model to serve as a robust tool for climate analysis and weather forecasting. We are confident that these endeavors will stimulate further advancements in the field of image recognition technology.

References

[1] Cheng Gong,Han Junwei & Lu Xiaoqiang.(2017).Remote Sensing Image Scene Classification: Benchmark and State of the Art.Proceedings of the IEEE(10),1865-1883.

[2] Illya Kokshenev & Antonio Padua Braga.(2010).An efficient multi-objective learning algorithm for RBF neural network.Neurocomputing(16),2799-2808.

[3] TesoFzBetoño Adrian,Zulueta Ekaitz,CabezasOlivenza Mireya,TesoFzBetoño Daniel & FernandezGamiz Unai.(2022).A Study of Learning Issues in Feedforward Neural Networks.Mathematics(17),3206-3206.

[4] Zachary Chase Lipton.(2015).A Critical Review of Recurrent Neural Networks for Sequence Learning..CoRR

[5] Hinton Geoffrey E,Osindero Simon & Teh Yee-Whye.(2006).A fast learning algorithm for deep belief nets..Neural computation(7),1527-54.

[6] Yue Ma,Xu Ji,Nasher M.BenHassan & Yi Luo.(2018).Automatic fault detection with Convolutional Neutral Networks.(eds.)CPS/SEG Beijing 2018 International Geophysical Conference Exposition Electronic papers(pp.808-812).Aramco Research Center-Beijing,Aramco Asia;EXPEC Advanced Research Center,Saudi Aramco;

[7] Andrew. G. Howard, Menglong Zhu, Bo. chen. Dmitry. Kalenichenko., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017, [Online]. Available: http:/ /arxiv.org/abs/1704.04861.

[8] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," Proceedings of the IEEE, vol. 105, no. 10, pp. 1865-1883, 2017.

[9] Rainio Oona,Teuho Jarmo & Klén Riku.(2024).Evaluation metrics and statistical tests for machine learning.Scientific Reports(1)

[10] Kaiming He,Xiangyu Zhang,Shaoqing Ren & Jian Sun 0001.(2015).Deep Residual Learning for Image Recognition..CoRR

[11] Karen Simonyan & Andrew Zisserman.(2014).Very Deep Convolutional Networks for Large-Scale Image Recognition..CoRR

[12] Alex Krizhevsky,Ilya Sutskever & Geoffrey E. Hinton.(2017).ImageNet classification with deep convolutional neural networks.Communications of the ACM(6),84-90.

[13] Y. Lecun, L. Eon Bottou, Y. Bengio, and P. H. Abstract|, “Gradient-Based Learning Applied to Document Recognition,” 1998.