Influence of Spectral Bands on Satellite Image Classification using Vision Transformers

Adithyan Sukumar, Arjun Anil, Sajith Variyar V. V, and Sowmya V[1], Moez Krichen[2], and Vinayakumar Ravi[3]

[1]Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India
[2]FCSIT, Al-Baha University, Al Baha, KSA
[2]ReDCAD, University of Sfax, Sfax, Tunisia
[3]Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia

v_sowmya@cb.amrita.edu

Abstract. Neural networks play an important role in satellite image classification. We know that the most common neural networks used in image classification tasks are convolutional neural networks (CNNs) [2]. In this paper, we explored the influence of the spectral bands in image classification using the Vision Transformer (ViT). Convolution is a local operation, and a convolution layer typically models only the relationships between neighborhood pixels. Transformer is a global operation, and a transformer layer can model the relationships between all pixels. This motivated us to use ViT for satellite image classification. Sentinel-2 EuroSAT image dataset, which consists of 27,000 images in ten classes, is used for the experiment. ViT model is trained with three band dataset, Red-Green-Blue (RGB) and compared with ViT model trained with RGB along with Near InfraRed (NIR) and with multispectral satellite image dataset (13 bands). Experimental results shows that NIR band combined with RGB was able to produce more accurate results comparing to RGB alone, whereas 13 band dataset outperformed both RGB and RGB & NIR datasets.

Keywords: Vision Transformer · Satellite Image Classification · Sentinel-2 Satellite Images · Deep Learning For Satellite Images

1 Introduction

Land cover refers to the surface on the ground that covers vegetation, infrastructure, water bodies, soil, etc. Land cover classification is important to monitor changes on the surface. Presently, image classification has become an important task of remote sensing for pattern analysis, and image analysis. Various researchers have developed and used different image classification methods. The classification of satellite images helps us to monitor various aspects such as pollution, forest cover mapping, wetland mapping, and land cover analysis, etc[8].

ViT was proposed by Alexey Dosovitskiy et al. [3] as a replacement for CNN in image classification tasks. ViT is used to classify datasets such as ImageNet, CIFAR-100, VTAB, etc., and was able to yield excellent results compared to CNN. ViT is now commonly used for remote sensing scene classification tasks for datasets such as Merced, AID, Optimal31, and NWPU [1]. Mark Pritt et al. [9] integrated metadata and image features from the IARPA fMoW dataset and used CNN as a set to produce good precision when classifying 63 different classes [5]. Leveraging the feature extraction capability of pretrained Resnet50 delivered promising results when trained on SAT4 dataset [7]. Jionghui Jiang et al. [6] used two CNN trained in RGB and NIR data, as normal CNN cannot fully exploit the information due to the high correlation between the bands. The normalized difference vegetation index can be calculated using NIR along with the red band. Using this as input to VGG, Alexnet, and Convnet helps the models to be efficient by giving high accuracy with fewer parameters that can be trained [12, 10]. Color to grayscale image conversion is one of the most important steps in image pre-processing. Sowmya et al. [11] used singular value decomposition based image conversion techniques to convert SIFT features combined with structure similarity index. The results show that the accuracy of the model increases with the conversion of images. In this paper, we use ViT to compare the RGB dataset with RGB & NIR and with all the 13 spectral band dataset. The objective is to show the influence of the addition of more bands into the dataset helps the ViT classify images more accurately. RGB bands along with the NIR bands were taken from the 13-band dataset to compare the RGB NIR dataset with the RGB datasets. 13-band dataset is taken to compare the multispectral dataset with the fewer-band data set. This paper is organized into sections as follows. Section 2 discusses the materials and methodology. Section 3 shows the results obtained provides information on the metrics used to evaluate the model and the values obtained followed by conclusion.

2 Methodology

ViT was trained and compared using three datasets (RGB, RGB & NIR, Multispectral, or 13 bands). For this paper, we used Sentinel-2 satellite images, which are freely accessible and provided by the Earth observation program Copernicus [4]. Dataset consists of a total of 13 bands, of which we require four bands for the study Red, Green, Blue, and Near-Infrared to make comparison between RGB and RGB & NIR. We use the 13 band dataset to compare both 3 band and 4 band dataset. Position of Red, Green, Blue, and NIR channels are one, two, three, and seven (considering the indices from zero) respectively. [4] as shown in Fig.2. The data used for the study are Sentinel-2 EuroSAT images which is divided into ten class labels, Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent crop, Residential, River, SeaLake. The dataset consist of 27000 .tif files which have all the bands of the spectrum as collected from the Sentinel-2 satellite. Each class label contains 2000-3000 images. A detailed breakdown of this dataset is given in Table 1.

Fig. 1. Representation of classes in the Sentinel-2 EuroSAT image dataset

Data augmentation is frequently used to help neural networks generalize better to unseen data and to increase the richness of the training data utilized. Various data augmentation techniques are applied to the dataset like resizing, normalization, random horizontal flip, random rotation, and random zoom. In order to maintain equality in comparison, augmentation is applied equally in all three datasets.

Table 1. A detailed description of the dataset

Class No.	Class Labels	Number of Images
1	*Annual Crop*	3000
2	*Forest*	3000
3	*Herbaceous Vegetation*	3000
4	*Highway*	2500
5	*Industrial*	2500
6	*Pasture*	2000
7	*Permanent Crop*	2500
8	*Residential*	3000
9	*River*	2500
10	*Sea Lake*	3000

The deep learning architecture used for image classification is Vision Transformer. ViT is a self-attention-based architecture, similar to the transformers we use in natural language processing (NLP) tasks. By the introduction of transformers in NLP was able to create a large success; here ViT is also such a transformer that was developed by making very few possible modifications. Usually, tokens are given as input for the NLP transformer, whereas in ViT, we split large images into patches along with linear embedding as the input. We train this model for image classification in a supervised method.

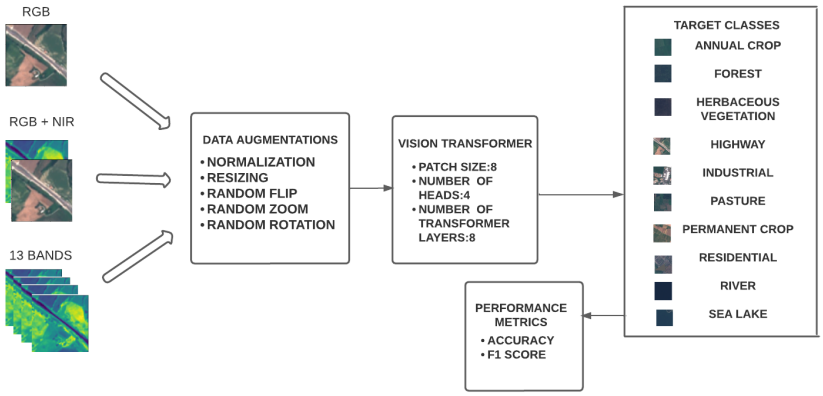

Fig. 2. Workflow for the satellite image classification using ViT.

The architecture consists of a patch and position embedding unit that splits input images into patches and gives positional embedding to those patches. These patches along with positional details are fed as input to the transformer encoder. The transformer encoder has alternating layers of multiheaded self-attention and MLP blocks. Layernorm (LN) is applied before every block and residual connections after every block. The last layer is a softmax layer as shown in Fig 3 [3].

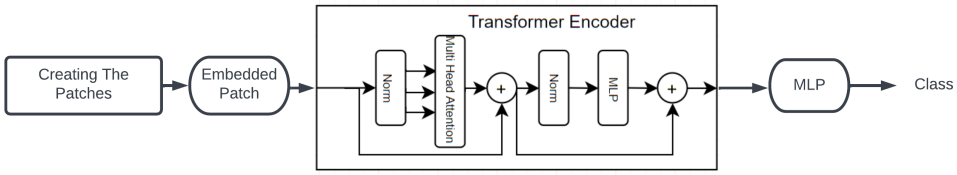

Fig. 3. Vision Transformer Archeitecture

To perform the experiment without bias, we used the same hyper-parameters for training ViT on these datasets. To make the model generalize, certain data augmentations like resizing, normalization, flipping, rotation, etc have been applied. The efficiency and accuracy of the classification of satellite images depend on the model selected for classification. ViT splits the images into a series of position-embedded patches, which are sent to the transformer encoder; By doing this, ViT is able to understand the local and global features of the image. The model was evaluated using accuracy, precision, recall, and the F1 score. These metrics serve as the basis to identify the errors during the classification process. To determine these metrics, testing data from each class are used.

3 Results and Discussions

The data augmentation applied were resizing, normalization, random horizontal flipping, random rotation, and random zoom. The original dataset had 27,000 images, but the experiment was conducted only for 20,000 images in order to maintain class imbalance problems. Out of the 20000 images, 70% of the data was taken for training, 10% of the data for validation and 20% for testing. Experimental setup was kept the same for all three experiments (RGB, RGB NIR, Multispectral). The experiment was carried out under the following circumstances in tensorflow, the learning rate was set at 0.001 with a batch size of 256. The image was broken down to 8 patches since it produced the best results via trial and error method. The number of transformer layers for ViT is taken as eight. From Fig.3, it can be inferred that 100 epochs give better validation

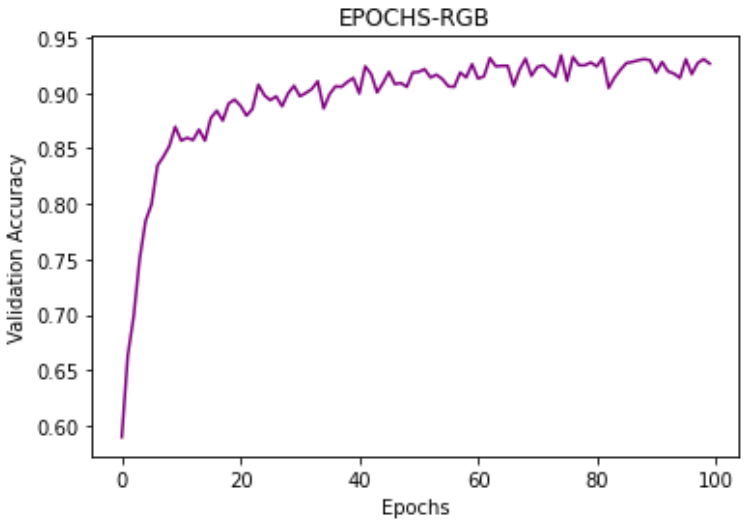

Fig. 4. Epochs Vs Validation Accuracy for RGB dataset

accuracy for RGB data. In order to study the effect of addition of NIR and multispectral bands experimental setup was kept same.

The metrics used to evaluate the performance of the models are accuracy, recall, precision, and f1-score. Accuracy, precision and F1-Score is computed using the equation.

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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

Table 2. Performance analysis of the ViT model for three datasets

ViT									
	RGB & NIR			RGB			Multispectral		
Classes	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Annual Crop	0.99	0.83	0.9	0.96	0.84	0.89	0.97	0.96	0.97
Forest	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.99	0.99
Herbaceous Vegetation	0.93	0.93	0.93	0.9	0.93	0.92	0.98	0.96	0.97
Highway	0.92	0.89	0.9	0.89	0.89	0.89	0.93	0.9	0.92
Industrial	0.96	0.95	0.96	0.97	0.93	0.95	0.95	0.98	0.96
Pasture	0.92	0.9	0.91	0.89	0.92	0.9	0.96	0.96	0.96
Permenant Crop	0.84	0.91	0.8	0.84	0.92	0.88	0.97	0.94	0.96
Residential Area	0.89	0.99	0.94	0.93	0.98	0.96	0.95	0.97	0.96
River	0.98	0.98	0.98	0.91	0.91	0.91	0.98	0.98	0.98
Sealake	0.99	0.99	0.99	0.98	0.96	0.97	1	1	1
Accuracy	0.939			0.929			0.955		

The ViT trained with the RGB dataset gave an accuracy of 0.929. The f1 scores for classes such as Sealake, Forest, and Residential Area were 0.97,0.98 ,0.96 respectively. From the classification report, it is understood that very few classes were classified with a precision, recall and F1 score greater than 0.95. Additionally, using the NIR band along with the RGB band for ViT gave an

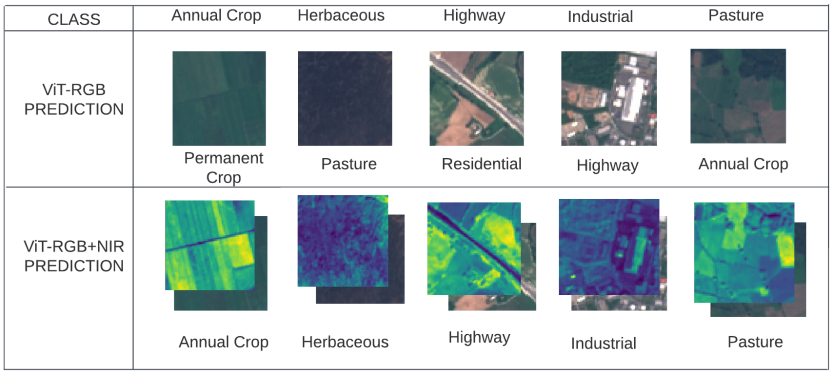

Fig. 5. Images predicted via model trained on RGB & NIR dataset which were misclassified by the model trained on the RGB dataset

accuracy of 0.939. Classes like Forest, River, SeaLake, Industrial and Residential

area was classified correctly in appreciable manner. In addition to this, the precision for annual crops, herbaceous vegetation, highway and pasture was high, and the recall for industrial was also high (Fig.4).

By using all the 13 bands provided in the dataset, the ViT gave an accuracy of 0.95. All classes except highway was predicted correctly with high precision, recall and F1 score. Most of the classes gave an F1 score greater than 0.96 and all SeaLake class test images were correctly classified (F1 score=1). Comparing the results in Table 2, the RGB dataset was able to classify only very few classes efficiently. The addition of the NIR band to the RGB dataset resulted in the increase of true positive values. Annual Crop, Herbaceous Vegetation, Highway, Industrial, Pasture, River and SeaLake are the classes were 4 band dataset outperformed 3 band dataset. Whereas Multispectral dataset on Vision transformer predicted most of the classes better than both 4 bands and 3 bands.

Table 3. Classwise comparison of RGB and RGB & NIR

Classes	Bands
Annual Crop	RGB & NIR
Forest	RGB
Herbaceous Vegetation	RGB & NIR
Highway	RGB & NIR
Industrial	RGB & NIR
Pasture	RGB & NIR
Permenant Crop	RGB
Residential Area	RGB
River	RGB & NIR
Sealake	RGB & NIR

The RGB bands contain information about rich color features of the image, and the NIR band contains details regarding sharp edges in the image.This shows that adding more bands to the dataset can provide more information, thus improving performance of machine learning models. Table 3 shows which bands perform better for each class, and for most classes RGB & NIR produced improved results. Capturing multispectral images requires more sensors and is expensive, compared to the requirement of RGB & NIR. Since RGB & NIR images are able to give better classification accuracy compared to RGB images and was also close to the accuracy of 13 band images, it is more cost efficient and computationally inexpensive.

4 Conclusions

In this paper, we have compared the performance of the vision transformer trained on RGB, RGB & NIR and multispectral data. Results show that addition of more bands increases ViT performance. ViT trained with RGB & NIR images

outperformed model trained with RGB images, whereas multispectral (13 bands) outperformed both. This shows the influence of addition of spectral bands on ViT. In the future works, we can improve the performance of the model using datasets with more bands and using the right combinations of bands. Apart from increasing the number of bands, the model performance can be further improved by increasing transformer layers and multilayer perceptron heads.

References

1. Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3):516, 2021.
2. Gaudenz Boesch. Vision transformers (vit) in image recognition – 2022 guide. https://viso.ai/deep-learning/vision-transformer-vit, 2022.
3. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
4. Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
5. Deepika Jaswal, Sowmya Vishvanathan, and Soman Kp. Image classification using convolutional neural networks. *International Journal of Scientific and Engineering Research*, 5(6):1661–1668, 2014.
6. Jionghui Jiang, Xi'an Feng, Fen Liu, Yingying Xu, and Hui Huang. Multi-spectral rgb-nir image classification using double-channel cnn. *IEEE Access*, 7:20607–20613, 2019.
7. Mohammed Abbas Kadhim and Mohammed Hamzah Abed. Convolutional neural network for satellite image classification. In *Asian Conference on Intelligent Information and Database Systems*, pages 165–178. Springer, 2019.
8. Gordana Kaplan and Uğur Avdan. Mapping and monitoring wetlands using sentinel-2 satellite imagery. 2017.
9. Mark Pritt and Gary Chern. Satellite image classification with deep learning. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2017.
10. T Tulasi Sasidhar, K Sreelakshmi, MT Vyshnav, V Sowmya, and KP Soman. Land cover satellite image classification using ndvi and simplecnn. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2019.
11. V. Sowmya, D. Govind, and K. P. Soman. Significance of contrast and structure features for an improved color image classification system. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 210–215, 2017.
12. Anju Unnikrishnan, V Sowmya, and KP Soman. Deep learning architectures for land cover classification using red and near-infrared satellite images. *Multimedia Tools and Applications*, 78(13):18379–18394, 2019.