DOI: 10.15849/IJASCA.220328.14

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 1, March 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Remote Sensing Image Classification Via Vision Transformer and Transfer Learning

Muhammad Saad Shahbaz khan, Muhammad Asim Rajwana

Department of Computer Science, National College of Business Administration & Economics Lahore, Sub Campus Multan,60000, Pakistan
Saadikhan335@gmail.com
Department of Computer Science, National College of Business Administration & Economics Lahore, Sub Campus Multan,60000, Pakistan
NCBAEMultan@yahoo.com

Abstract

Aerial scene classification, which aims to automatically tag an aerial image with a specific semantic category, is a fundamental problem for understanding high-resolution remote sensing imagery. The classification of remote sensing image scenes can provide significant value, from forest fire monitoring to land use and land cover classification. From the first aerial photographs of the early 20th century to today's satellite imagery, the amount of remote sensing data has increased geometrically with higher resolution. The need to analyze this modern digital data has motivated research to accelerate the classification of remotely sensed images. Fortunately, the computer vision community has made great strides in classifying natural images. Transformers first applied to the field of natural language processing, is a type of deep neural network mainly based on the self-attention mechanism. Thanks to its strong representation capabilities, researchers are looking at ways to apply transformers to computer vision tasks. In a variety of visual benchmarks, transformer-based models perform similar to or better than other types of networks such as convolutional and recurrent networks. Given its high performance and less need for vision-specific inductive bias, the transformer is receiving more and more attention from the computer vision community. In this paper, we provide a systematic review of the Transfer Learning and Transformer techniques for scene classification using AID datasets. Both approaches give an accuracy of 80% and 84%, for the AID dataset.

Keywords: *remote sensing, vision transformers, transfer learning, classification accuracy.*

Received 7 December 2021; Accepted 3 March 2022

1 Introduction

Remote detection images are an important source of earth observational data, which can help us measure and examine deep structures on the surface of the earth. Thanks to advances in Earth's diagnostic technology, the number of images analyzed remotely increases dramatically. This has provided a particularly urgent way to find ways to use the growing images of remote sensors for global psychiatry. Therefore, it is very important to understand the big and complex images of distant sensations. As a major problem and challenge for better interpretation of remote sensor images, the classification of remote sensor imagery has become the subject of applied research. The classification of remote analyzes includes the precise labeling of accompanying images of remote sensors and semantic categories identified earlier, as shown in Figure 1. Over the past few decades, extensive research has been done on the classification of remote events. depending on the situation. -common use, such as urban planning, natural hazard detection, environmental monitoring, plant mapping, and geographical object recognition.

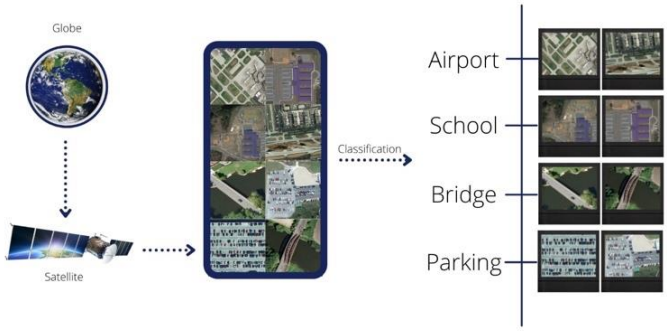


Figure 1: Satellite Image Classification

With the improvement of the atmospheric resolution of remote sensing images, the remote sensing classification forms three branches of classification at different levels: pixel scale, object level, and area-level classification. It is important to note here that we use the term “Remote sensing image classification” as a general concept that includes the pixel scale, object level, and the general classification of the scene. In particular, researchers in early literature aimed to classify sensory images at the pixel or sub-pixel level by assigning each pixel to the scene classification and semantic class. In 2001, Blaschke and Strobl [1] questioned the concept of pixel research and came to the conclusion that the analysis of remote sensing images at the object level is better than pixel analysis. They suggested that researchers focus on object-level analysis aimed at identifying objects in remote sensing images. where the word "object" refers to semantic objects or surface units. Later, in the last two decades, several methods of analyzing remote sensing images on an object scale dominated the analysis of remote sensing images. The remarkable success of special land use tags is complemented by pixel scales and scene classification. Because of the high resolution of remote sensing images, remote sensing images can have different classes and different objects. Pixel scale

methods may not be sufficient to accurately classify each time. In these situations, it is very interesting to understand the importance of worldly content and images of remote sensing. Recently, a new concept of scene-level analysis of remote sensing images at a stage level has been proposed. Remote scale analysis image classification attempts to classify each remote sensing image patch (256x256) in a semantic class, as shown in Figure 1. The "scene" of an object represents a part of an image cut from a large remote sensing image that contains clear semantic information on the earth's surface. This is an important step as it can provide visual and distinctive information in almost any computer vision activity. After extensive research, various methods of classification of remote images have emerged. The number of journals in the classification of remote images increased significantly after 2014 and 2017. There are two reasons for the increase. On the one hand, the deep learning technology for remote sensing data analysis began around 2014. On the other hand, the characteristics of the remote sensing classification in 2017 are more clearly visible, which is conducive to the development of deep learning-based remote sensing image scene classification. Self-attention-based, especially Transformers Vaswani et al. [2] have become a model for choosing a natural language processing (NLP). The main method is to train a large text dataset and then fine-tune a small amount of specific data. Thanks to the calculation and conversion efficiency of the Transformer, it has been possible to teach models of non-standard sizes with parameters exceeding 100B. With the growth of models and data sets, there are still no signs of general performance. However, in computer vision, convolutional architectures are still common LeCun et al. Krizhevsky et al., He et al. [3,4] Inspired by the success of NLP, some researchers tried to combine architectures such as CNN with self-actualization Wang et al. Carion et al. [4] while others completely changed the tremor Ramachandran et al. Wang Et al. [5] Although these latter models are theoretically related, they have not been effectively scaled to modern hardware upgrades due to the use of specialized focusing systems. Therefore, for better image recognition, the standard architecture of ResNet is still very high Mahajan et al., Xie et al. & Kolesnikov et al. [6] Encouraged by the success of the Transformer addition in NLP, we tried to use the standard Transformer directly on images with as few adjustments as possible. To this end, we split the image into patches and provide the layout of these patches as input to the Transformer. In NLP applications, image patches are treated in the same way as tags (words). We teach the image classification model in a controlled way. In this study, we

ensemble two different techniques: Vision Transformer (VIT) and Transfer Learning (TL) are used to test the AID dataset. The rest of the paper is organized as follows: Section 2 discusses some related work. Section 3 discusses the Aerial Image dataset that will be used for the experiment. The experiment methodology and the fundamental concept of the two ensemble techniques being investigated are discussed in Section 4. Experiment results and discussions are provided in Section 5. Finally, Section 6 contains the conclusion for this work.

2 Related Work

Remote Sensing images, regardless of environmental resolution, are a reflection of the earth's surface Hofmannrt al. [6] and an important asset for the ability to record multiple-scale information within an area. Depending on the type of information required, pixel base, object base, or basic building blocks can be provided. However, no effective and comprehensive strategy has been reported better integrate these services due to a better data connection. On the other hand, DL can be representative. and plan several levels of information to explain the complex relationships between data Hofmann et al. In fact, DL technicians may have different levels Take a picture and stir Liu, et al. Take the understanding of the area as an example using DL curtains can be seen as Integrate change with different local conditions and development systems inherited from small projects without a department One step or step is required. Despite its great potential, DL cannot be used directly many RS features with many limitations Tape. Some RS images, especially light spectrum images, have hundreds of bands that can lead to the formation of small particles. A very large cube corresponding to a large number in the network prepared by the arteries Camps-Valls et al. On the website, Important information is also the geometric pattern of each band, band vectors. But how to use this information is still necessary for further research. There are still issues with high-resolution images in RS resolution that only have a green, red, and blue channel, similar to the DL class. In practice, there are several systems that can be developed to create a dedicated network. In addition, images taken by different sensors show clear differences. How to transfer a pre-installed network to other images is not yet known.

Transformers have been used by Vaswani et al. [2] It is used for machine translation and has become the most advanced method in many NLP works. The main types of Transformer fundamentals are usually taught early in a large collection and then well prepared for the existing work: BERT Devlin et al. [7] is used to do pre-training supervised tasks, during the GPT task line Language uses formulation as a function. Pre-training tasks Radford et al.& Brown et al. [9]

A naive application of self-attention to images would require each pixel to pay attention to every other pixel. Due to the secondary cost of the number of pixels, it cannot be extended to the actual input size. Therefore, when applying Transformer in the context of image processing, several attempts have been made to estimate in the past: Parmar et al. [9] apply self-attention only to the local neighborhood of each query pixel, rather than globally. This multiheaded doll product local self-attention block can completely replace convolution Ramachandran et al., Cordonnier et al., Zhao et al. [10,11,12] Or Sparse Transformers Child et al. and other works adopt a scalable global self-attention approach to be adequate for image scaling. Another method of attention is to apply it to blocks of different sizes Weissenborn et al. [13], only in extreme cases along a single axis Ho et al. &, Wang et al. [14,15]. These dedicated architects have shown promising results in computer vision tasks, but they need to effectively implement complex engineering in hardware accelerators.

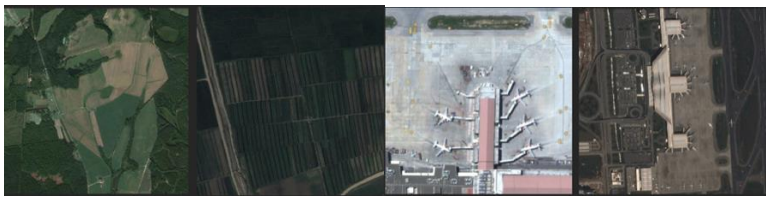
Combining convolutional neural networks (CNN) with self-attention forms has also attracted a lot of interest, for example, adding resource maps for classification imaging Bello et al. [16] using self to further process the output of CNN. Note for example, for object detection Hu et al., Carion et al. [17], video processing Wang et al. Sun et al. [18], image classification Wu et al. unsupervised object discovery Locatello et al. [19], or unified text view task Chen et al., Lu et.[20]

3 Dataset

This section provides some detailed information about the data sets we used in our experiments and the number of samples in each data set. We use a ratio of 70% training set,10% validation set, and 20% test set.

3.1 AID: Aerial Image Dataset

AID is a new special image file that can collect images from Google Earth. Note that although the images of Google Earth are in RGB, compared to the original aerial images, it turns out that there is no significant difference between Google Earth and the original aerial images, even at the raster level. This means that images from Google Earth can also be used as aerial imagery to evaluate visual classification algorithms.
The new dataset contains 30 types of aerial photos. In total, the AID dataset contains 10,000 images in 30 categories. In addition, all sample images of each type of AIDS have been carefully selected from different countries and regions around the world (especially China, USA, UK, France, Italy, Japan, Aponia, Germany, etc.). and they are gathered at different times and seasons under different recording conditions, increasing the variety of data. Some examples from different classes are shown in figure 2.



Airport Farmland

Railway Station Church

Mountain Commercial

Figure 2: AID Dataset Samples

We take a ratio of 70% training set, 20% test set and 10% validation set per class as shown in Table 1.

Table 1: There are 30 classes and 10,000 images

Classes	Training	Validation	Testing	Total	Classes	Training	Validation	Testing	Total
Airport	252	36	72	360	**Parking**	273	39	78	390
Bare land	217	31	62	310	**Playground**	259	37	74	370
Baseball field	154	22	44	220	**Pond**	294	42	84	420
Beach	280	40	80	400	**Port**	2662	38	76	380
Bridge	252	36	72	360	**Railway Station**	182	26	52	260
Center	182	26	52	260	**Resort**	203	29	58	290
Church	168	24	48	240	**River**	287	41	82	410
Commercial	245	35	70	350	**School**	210	30	60	300
Dense residential	287	41	82	410	**Sparse Residential**	210	30	60	300
Desert	210	30	60	300	**Square**	231	33	66	330
Farmland	259	37	74	370	**Stadium**	203	29	58	290
Forest	175	25	50	250	**Storage Tanks**	252	36	72	360
Industrial	273	39	78	390	**Viaduct**	294	42	84	420
Meadow	196	28	56	280	**Medium residential**	203	29	58	290

Park	245	35	70	350	**Mountain**	238	34	68	340

3.2 Data Augmentation

Data augmentation is a simple but effective tool that can increase the size and diversity of data sets. For data that cannot reach a large number of labeled data, this is a basic step. Cubuk, ED, and others. [21] Data enhancement uses a variety of surgical techniques to provide additional training samples from existing samples while maintaining the efficiency of the original class assignment. Training the model on augmented data helps solve the problem of overhaul, which improves the durability of the model and the overall capacity. Typical data processing techniques create new samples using simple geometric transformations (such as rotating, increasing size, subtracting, translation, and rotating or combining them), as shown in Figure 3.

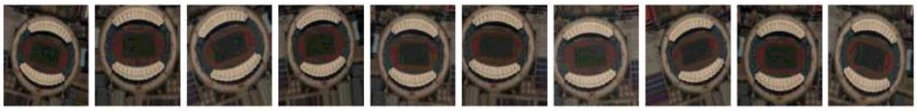


Figure 3: Data Augmentation of Stadium Images

4 The Proposed Method

Let, a high-resolution image resize into 224 x 224 images. Then we apply data augmentation techniques. Unless we have a large image dataset, it is recommended that we make appropriate but realistic changes to our training images, such as the transformation of scaling, cropping, and horizontal flipping to determine the training dataset. This is helpful to explore the training data and reduce overfitting. After data augmentation, we train our model and test on testing data and predicate the class. In figure 4 flow diagram show all stages done to achieve the classification.

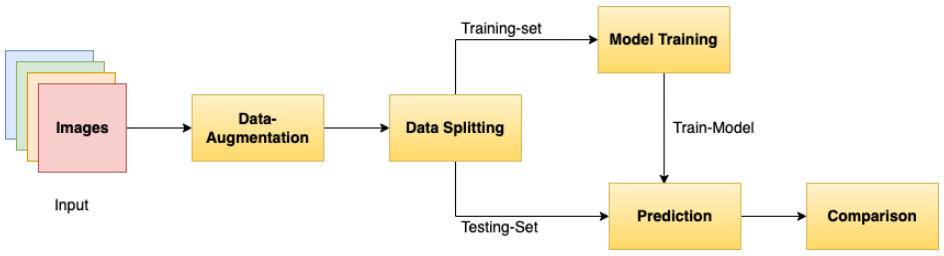


Figure 4: Flow Diagram of Methodology

4.1 Vision Transformers

When it comes to image classification the first model you can think of is definitely a Convolutional neural network, Resnet is the best among CNN models.

Resnet was the best solution to image classification. Vision transformer and VIT is a new State of the art technique in image classification. Vit beats Resnet by a small margin. Vit has been pre-trained on a sufficiently large dataset the bigger the data set the greater the advantage of the Vit over ResNet. Transformers was originally developed in 2017 for natural language processing, Vit is a successful application of transformer in computer vision but the model of vit does not have novelty, Vit is exactly the encoder network of the transformer as shown in figure 5.

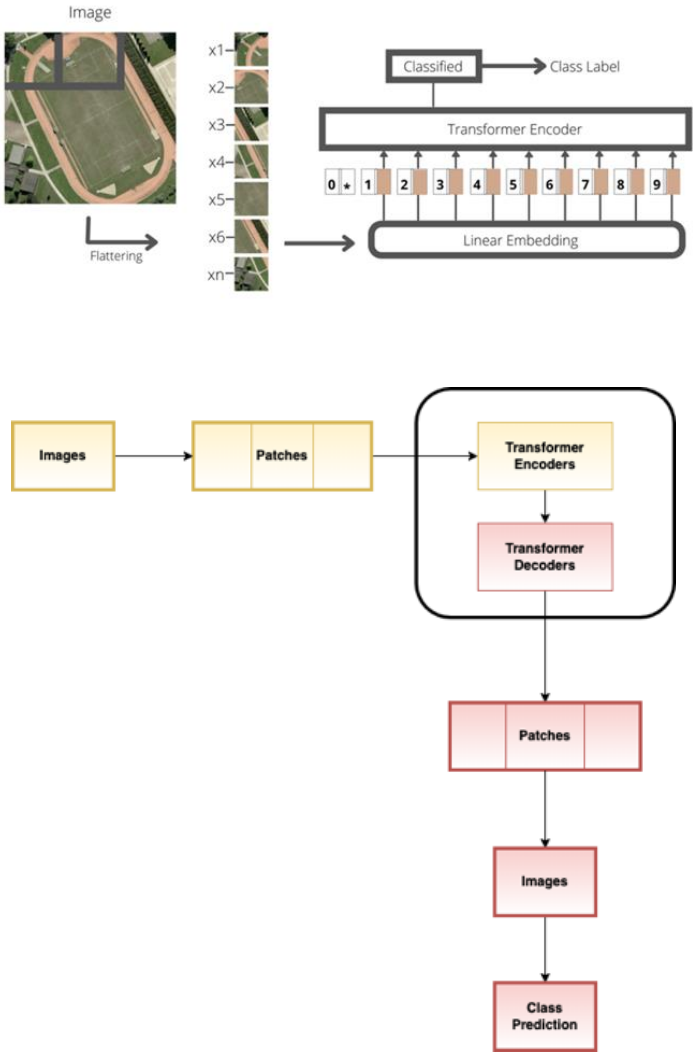


Figure 5: Vision Transformer Model and Flow Diagram

4.2 Positional Encoding

Look at the given image below, we portion the image into 9 nine patches here is a copy of the image look at the image on the right we change the position of the patches. Now both images are different however changing the Z vector is not affecting the final output because the z vector is not containing the position encoding of the patches so both images will be the same. But transformer

perspective this is unreasonable because the images on the left and the right are different. We hope the transformer knows that these two images are different so we assign the positional information to the patches and add the positional encoding to the Z vectors in this way if the two patches are swapped their positional encoding will change and therefore the output of the transformer will be different.

Let's come back to the new network, where building the X1 to Xn vectorization of the patches Z1 to Zn is the resulting output of the linear transformation and passional encoding they are the representation of the n patches they capture both the content and position of the patches aside from the n patches we use CLS token for the classification and embedding layer takes as input the CLS token and outputs vector Z0.Z0 has the same shape as the other Z vectors. We use the CLs token because the output of the transformer in the position will be used for the classification. The sequence of the Z0 to Zn to multi-head self-attention layer the out of the these HAS layers is N+1 vectors, then apply a dense layer. The output of the dense layer is also an n+1 vector. Then add and multi-head self-attention layer and dense layer. We can add many self-attention layers and dense layers as we required. This is also called a transformer encoder network. Vectors C0 to Cn are the output of the transformer. For the classification task, we do not need the C1 to Cn vectors so we ignore them what we need is a C0 vector. It is a feature vector extracted from the image; the classification is based on the C0. Feed the C0 vector to the SoftMax classifier. The classifier outputs vector P. Shape of P is a number of classes if the dataset has 30 classes, then P has 30 dimensional. During the training, we compute the cross-entropy of the vectors P and the ground truth then compute the gradient of the cross-entropy loss with respect to the model parameters and perform gradient descent to update the parameters.

4.3 Transfer Learning

The base model is from the Visual Geometry Group from Oxford's which is also called VGG. Here we are using the VGG19 model. This is the pre-trained model on the ImageNet dataset, ImageNet is a large dataset consisting of 1.4M images and 1000 classes. ImageNet is a collection of different categories; this knowledge helps to classify satellite images in a particular data. First, use the pre-train VGG19 model with weights that are trained on ImageNet. By doing 'include_top = False' upload the network without the top layers of classification. This makes it easier to extract features.

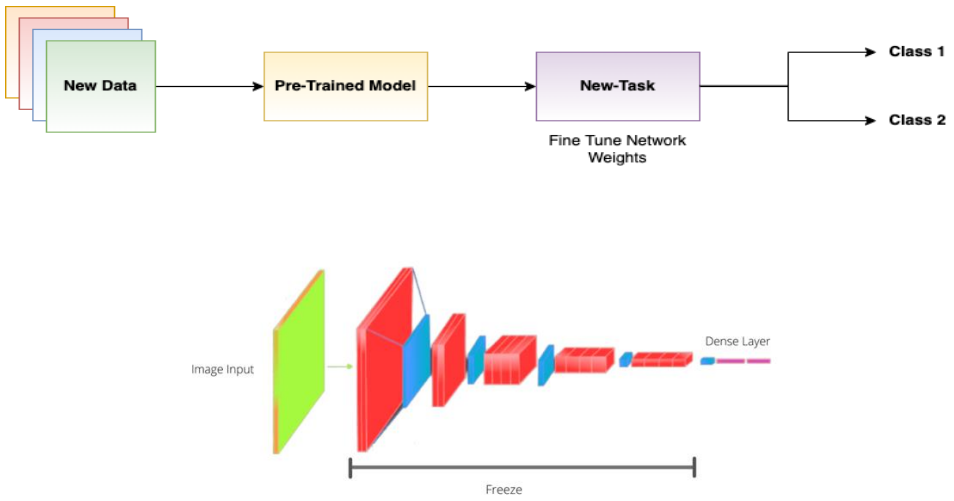


Figure 6: Graphical Representation of Model and Flow Diagram

In this step, we will freeze the top base layer which we discuss above, and use it as a feature extractor. The architecture of the model and flow diagram of transfer learning is shown figure 6. It is important to freeze the convolutional base before assembling and learning the model. Freezing the "layer. trainable = False" prevents weight in the specified column to be updated during training. VGG19 has many layers, so setting up a complete trainable layer false which means all the trainable flag is frozen.

4.4 Model Training

To better understand the implications of the different methods and techniques used for remote sensing data classification, we performed two main experiments using the two different models in our thesis. The first one is the Transfer learning & CNN model. The Second model is Transformers and also compares the results of both models with randomly placed weights. The hyperparameter option has a significant impact on CNN's performance. But our main goal here is to investigate the results of transfer learning and also increase performance/accuracy. The model is trained by using Keras and TensorFlow as backend. The table 2 represents the training parameters which we used in during experiments.

Table 2: Training Parameters

Batch size	Optimizer's	Epochs
32	SGD	10
32	Adam	30
32	Adagrad	60

32	Adamax	80
100	Adam	100
40	SGD	200
500	Adamax	100
500	Adam	500

5 Results and Discussions

5.1 Performance of the Vision Transformer model

Firstly, classification results (%) with augmentation are shown the highest accuracy on epochs of 500. The vision transformers result without augmentation results show that Vision transformers need a higher amount of data for training. That's why without augmentation result accuracy is quite low. The results are clearly shown in both figure 7 and table 3 that the transformers need a large size of data to train the model for better results.

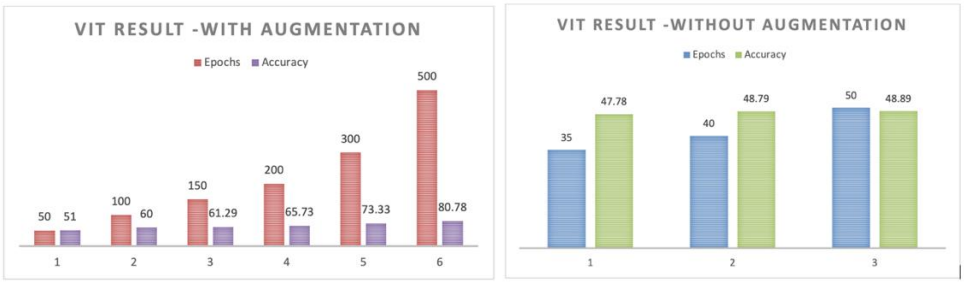


Figure 7: VIT Result With & Without Augmentation

Table 3: Vision Transformers Accuracy

No.	*Vision Transformers*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1		0.43	0.92	0.55	0.60
2		0.48	0.67	0.60	0.63
3		0.80	0.87	0.87	0.89

5.2 Performance of Transfer Learning Model

Here we discuss the performance of transfer learning on aerial image datasets. The figure 7 shows the left graph of transfer learning with Adamax optimizer which gives 84% accuracy of the model on remote sensing image classification. This 84% accuracy achieve on parameters of batch size 32 and 100 epochs. This graph represents that after 100 epochs when we increase the epoch size the performance of the model decrease. The right graph of figure 8 shows the other optimizer of

Adagrad which shows maximum accuracy of 77.7%. These experiments show that the when we increase the number of epochs the performance is also increased.

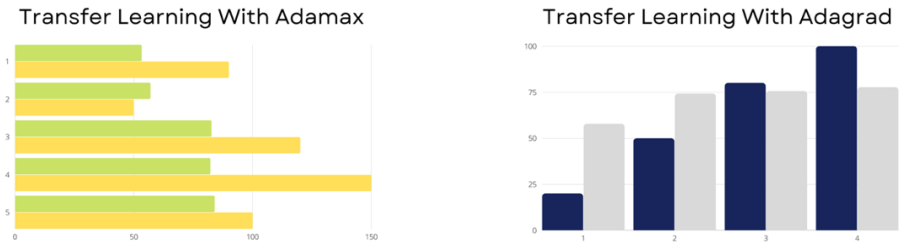


Figure 8: Performance of Transfer Learning

Lastly, we discuss about the comparison of multiple optimizers performance. The 1st is Adamax 2nd is Adagrad and 3rd is SGD. Three of them quite well on remote sensing images classification. The SGD and Adamax performance almost near to each other with 83.4% and 84% accuracy as shown in figure 9. Adagrad performance increase by increasing the epoch size in training time, same as Adamax there accuracy is also increased by increasing the number of the epochs but at some point, when the performance increasing rate goes vary slowly like accuracy increase in points. Table 4 represent the accuracy, precision, recall and F1-scoreTransfer Learning.

Figure 9: Comparison of Optimizers Performance

Table 4: Transfer Learning Accuracy

No.	*Transfer Learning*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1		0.77	0.82	0.93	0.88
2		0.83	0.83	1.0	0.91
3		0.84	0.88	0.95	0.91

5.3 Discussion

We further investigate that apply different batch size effects the performance of the propose method. When give bigger batch size to the model it learns faster as

compare to the smaller batch size. But bigger batch size required high resources. Plus, if give bigger batch size to the model there is chance the model will degrade the performance. For example, in our case when we apply batch size of 100 it gives 65.73% accuracy as compare to when I give bigger batch size like 500 it gives 61.29% accuracy. Model degrade the performance almost like 5%.

6 Conclusion

In this study, we propose a method for the classification of remote sensing images, based on vision transformers and transfer learning. In contrast to CNN, the vision transformer model can capture long-term dependencies between patches via the attention module. The proposed method was evaluated using public remote sensing image data set, and the experimental results demonstrated the effectiveness of these new networks in improving classification accuracy compared to the most modern methods. In addition, we show that using a combination of data expansion techniques can help further improve classification accuracy. We also observe that for transformers we need a big dataset for better results because a large dataset in size works better on transformers technique but it requires higher computation cost or resources. We apply another technique of transfer learning on remote sensing images-based classification, our proposed technique gives slightly better results than the state-of-the-art paper. Our transformer technique results are 80% on the other hand transfer learning results show 84% accuracy on the Aerial Image Dataset. The presented technique reports stable forecast model generation. Although we have a limited resource of GPU and RAM memory, our proposed method performs much better with respect to the hardware.

References

[1]. Blaschke, Thomas, and Josef Strobl. "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS." *Zeitschrift für Geoinformationssysteme* (2001): 12-17.

[2]. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[3]. LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1.4 (1989): 541-551.

[4]. Ghali, Rafik, et al. "Wildfire Segmentation Using Deep Vision Transformers." *Remote Sensing* 13.17 (2021): 3527.

[5]. Wang, Teng, et al. "End-to-End Dense Video Captioning with Parallel Decoding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[6]. Cao, Hu, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." *arXiv preprint arXiv:2105.05537* (2021).

[6]. Gu, Xiaolin, et al. "Classification-IoU Joint Label Assignment for End-to-End Object Detection." *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, Cham, 2021.

[7]. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[8]. Chada, Rakesh, et al. "Error Detection in Large-Scale Natural Language Understanding Systems Using Transformer Models." arXiv preprint arXiv:2109.01754 (2021).

[9]. Chen, Hao, Zipeng Qi, and Zhenwei Shi. "Remote Sensing Image Change Detection With Transformers." IEEE Transactions on Geoscience and Remote Sensing (2021).

[10]. Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." arXiv preprint arXiv:1906.05909 (2019).

[11]. Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "On the relationship between self-attention and convolutional layers." arXiv preprint arXiv:1911.03584 (2019).

[12]. Zhou, Luowei, et al. "End-to-end dense video captioning with masked transformer." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[13]. Weissenborn, Dirk, Oscar Täckström, and Jakob Uszkoreit. "Scaling autoregressive video models." arXiv preprint arXiv:1906.02634 (2019).

[14]. Wu, Xiaolei, et al. "StyleFormer: Real-Time Arbitrary Style Transfer via Parametric Style Composition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[15]. Wu, Bichen, et al. "Visual transformers: Token-based image representation and processing for computer vision." arXiv preprint arXiv:2006.03677 (2020).

[16]. Bello, Irwan, et al. "Revisiting resnets: Improved training and scaling strategies." arXiv preprint arXiv:2103.07579 (2021).

[17]. Cao, Hu, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." arXiv preprint arXiv:2105.05537 (2021).

[18]. Wang, Yuqing, et al. "End-to-end video instance segmentation with transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[19]. Locatello, Francesco, et al. "Object-centric learning with slot attention." arXiv preprint arXiv:2006.15055 (2020).

[20]. Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." European conference on computer vision. Springer, Cham, 2020.

[21]. Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation strategies from data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.