Satellite Image Analysis using Vision Transformers: Achieving SOTA in Vision Classification

Manav Garg[1], Atharva Subhash Bodke[1], Dr. E. Sudheer Kumar[2*]

[1]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Chennai, India
[2*]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Chennai, India
[*]Author to whom correspondence should be addressed.
Email: sudheerkumar.e@vit.ac.in

***Abstract*—Satellite imagery is used for different purposes including environmental monitoring, urban planning, and disaster response, among others. However, analyzing these images accurately and efficiently remains a challenge due to their large-scale and complex nature. In this study, we address this challenge by employing Vision Transformers (ViTs) to achieve state-of-the-art (SOTA) performance in satellite image analysis. With this in mind, we will explore the possibility of applying ViTs to a direct analysis of satellite images, recording the global relation dependencies and making land cover and land use classifications more accurate. The thought process here further leverages on recent development in computer vision particularly the success of ViTs in image recognition applications. Our task is to fine-tune pre-trained ViT models on the EuroSAT dataset, which is satellite images annotated with land cover classes. EuroSAT has 10 classes of land covers. By optimizing ViTs' performance on EuroSAT, we aim to surpass traditional methods in satellite image classification, offering more accurate insights into Earth's surface dynamics. This research contributes to advancing satellite image analysis and enhances tools for environmental monitoring, urban planning, and land management. We use evaluation technique based on metrics like classification accuracy that proves the efficacy of the suggested approach with reference to existent methods. This research also compares two Vision Transformer models, the base ViT and the CCT (Compact Convolutional Transformer) model.**

***Index Terms*—Vision Transformers, PyTorch, Satellite Image Analysis, EuroSAT**

I. Introduction

Computer vision, a field at the intersection of artificial intelligence and image processing, has witnessed remarkable advancements in recent years. Convolutional Neural Networks (CNNs) have been the cornerstone of many breakthroughs in computer vision tasks such as image classification, object detection, and segmentation. However, the traditional architecture of CNNs, with its reliance on convolutional layers, may not fully capture long-range dependencies in images, especially in scenarios where global context is crucial for accurate analysis.

In this context, Vision Transformers (ViTs) have emerged as a novel architecture that challenges the dominance of CNNs in computer vision tasks. ViTs leverage the transformer architecture, originally proposed for natural language processing tasks, to process images directly as sequences of patches. This approach enables ViTs to capture global dependencies in images through self-attention mechanisms, allowing them to analyze the context of an entire image more effectively. A study from Oct.2023 showed that the reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train. [3]

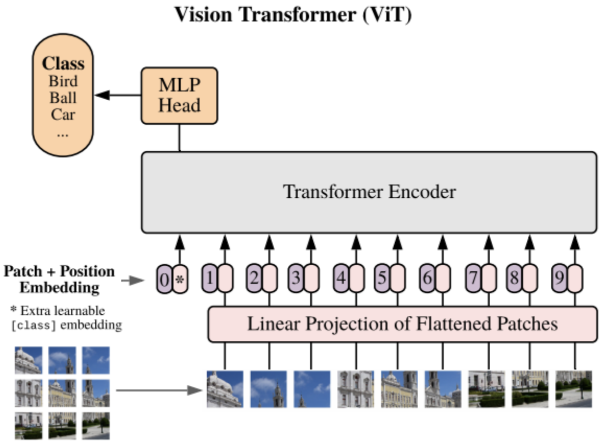

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Fig. 1: Vision Transformer Model Overview

Despite their promising potential, ViTs are relatively new to the field of computer vision, and their effectiveness compared to CNNs in various tasks is still under exploration. In this study, we delve into the application of ViTs for the automated classification of land use. By fine-tuning pre-trained ViT models on annotated satellite image datasets, we aim to evaluate the performance of ViTs in comparison to traditional CNN-based approaches.

A paper authored by Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, Humphrey Shi [4] showed how researchers have come to believe that because of corresponding growth in parameter size and amounts of

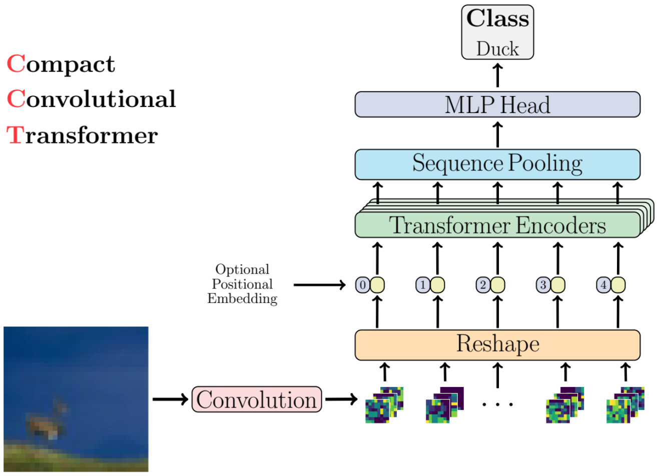


Fig. 2: CCT Model Overview

training data, transformers are not suitable for small sets of data. This trend leads to concerns such as: limited availability of data in certain scientific domains and the exclusion of those with limited resource from research in the field. The mentioned paper show for the first time that with the right size, convolutional tokenization, transformers can avoid overfitting and outperform state-of-the-art CNNs on small datasets. The models are flexible in terms of model size, and can have as little as 0.28M parameters while achieving competitive results.

Through extensive experimentation and analysis, we seek to assess the strengths and limitations of ViTs in medical image classification tasks. Our study aims to contribute to the growing body of research on ViTs and provide insights into their potential as a valuable tool for improving diagnostic accuracy in clinical settings.

II. Related Work

Several prior studies have explored the domain of image segmentation and object detection as well using Vision Transformers which marks notable achievements. A notable work by Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles and Hervé Jégou [7] proposed a competitive convolution-free transformer by training on Imagenet only. They trained them on a single computer in less than 3 days. Their reference vision transformer (86M parameters) achieves top-1 accuracy of 83.1% (single-crop evaluation) on ImageNet with no external data. More importantly, they introduced a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. We show the interest of this token-based distillation, especially when using a convnet as a teacher. This leads to the report results competitive with convnets for both Imagenet (where we obtain up to 85.2% accuracy) and when transferring to other tasks. [7]

One study showed that by using transfer learning and fine-tuning with RGB bands, we can achieve an impressive 99.19% accuracy in land use analysis. Such findings can be used to inform conservation and urban planning policies. [5]

Recently, there has been a growing interest in exploring Transformers for vision tasks, exemplified by the Vision Transformer (ViT) model for image classification. However, it has been observed that ViT underperforms compared to Convolutional Neural Networks (CNNs) when trained from scratch on midsize datasets like ImageNet. This is attributed to two main factors: firstly, the simplistic tokenization of input images fails to adequately capture important local structures such as edges and lines among neighboring pixels, resulting in low training sample efficiency; secondly, the redundant attention backbone design of ViT limits feature richness within fixed computation budgets and training samples.

To address these limitations, a new approach called Tokens-To-Token Vision Transformer (T2T-ViT) has been proposed. T2T-ViT incorporates two key innovations: firstly, a layer-wise Tokens-to-Token (T2T) transformation is introduced to progressively structure the image into tokens by recursively aggregating neighboring tokens into one, thereby enabling the modeling of local structures represented by surrounding tokens and reducing token length. Secondly, an efficient backbone with a deep-narrow structure for vision transformers is devised, drawing inspiration from CNN architecture design following empirical study.

Notably, T2T-ViT achieves a reduction in parameter count and Multiply-Accumulate operations (MACs) by half compared to vanilla ViT while yielding a more than 3.0% improvement when trained from scratch on ImageNet. Furthermore, T2T-ViT outperforms ResNets and achieves comparable performance with MobileNets when directly trained on ImageNet. For instance, a T2T-ViT model with a comparable size to ResNet50 (21.5 million parameters) can attain 83.3% top-1 accuracy with an image resolution of 384×384 on ImageNet. [8]

In the domain of remote sensing, Semi-supervised Learning (SSL) faces challenges due to limited labeled data availability and class imbalance in datasets. This study proposes a solution by generating synthetic labels and employing iterative class redistribution through resampling. Evaluation on EuroSAT, UCM, and WHU-RS19 datasets shows superior performance compared to previous methods. On balanced UCM, it outperforms MSMatch and FixMatch by 1.21% and 0.6%, respectively, while on imbalanced EuroSAT, it surpasses them by 1.08% and 1%, respectively. This approach reduces the need for labeled data, consistently outperforms alternatives, and addresses model bias caused by class imbalance. [6]

Another study proposes the use of generative models (GANs) for augmenting the EuroSAT dataset for the Land Use and Land Cover (LULC) Classification task. It used DCGAN and WGAN-GP to generate images for each class in the dataset. The authors then explored the effect of augmenting the original dataset by about 10% in each case on model performance. The choice of GAN architecture seems to have no apparent effect on the model performance. However, a combination of geometric augmentation and GAN-generated images improved baseline results. The study shows that GANs augmentation can improve the generalizability of deep classification models on satellite images. [1]

Deep learning techniques have long been effective in addressing remote sensing challenges. CNN-based models, particularly, excel in land classification tasks using satellite or aerial images, albeit often requiring substantial memory resources. Conversely, there's a growing demand for smaller models suitable for applications like unmanned aerial vehicles, which operate under memory constraints. However, downsizing CNN models typically sacrifices accuracy.

Experimental results demonstrate a significant enhancement in land classification accuracy, particularly for small-sized CNN models, validating the effectiveness of the proposed method. [2]

III. Dataset

A. EuroSAT

A novel dataset based on Sentinel-2 satellite images covering 13 spectral bands and consisting out of 10 classes with in total 27,000 labeled and geo-referenced images. Check Figure 3 for reference.

IV. Methodology

The methodology encompasses several key steps to effectively train a ViT (Vision Transformer) model for land classification using the EuroSAT dataset.

A. Dataset Loading and Preprocessing

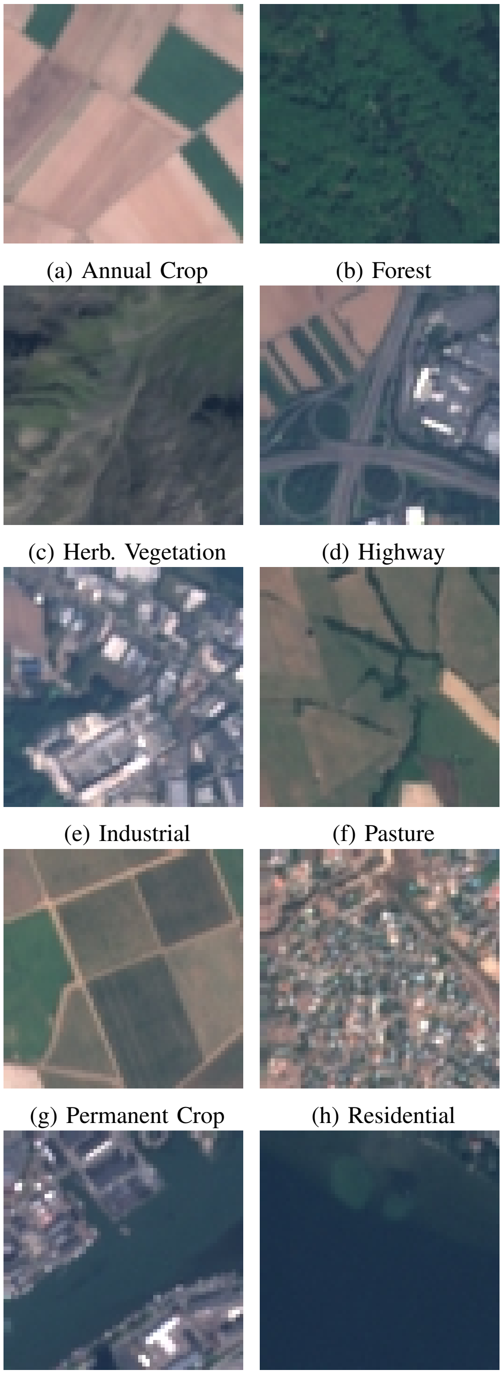
The EuroSAT dataset is loaded from the specified directory using the `load_eurosat_dataset` function. Images are resized to a uniform size of 128×128 pixels and normalized to have a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$.

B. Dataset Splitting

The dataset is split into training, validation, and test sets using the `random_split` function from PyTorch. The sizes of the splits are defined as 80%, 10%, and 10% of the total dataset size, respectively.

C. Vision Transformer Model Setup

```
# Base Vision Transformer
v = ViT(
    image_size = 128,
    channels = 3,
    patch_size = 32,
    num_classes = 10,
    dim = 1024,
    depth = 6,
    heads = 16,
```



(a) Annual Crop

(b) Forest

(c) Herb. Vegetation

(d) Highway

(e) Industrial

(f) Pasture

(g) Permanent Crop

(h) Residential

(i) River

(j) Sea Lake

Fig. 3: Sample Images from the EuroSAT dataset and their respective classes

```
    mlp_dim = 2048,
    dropout = 0.1,
    emb_dropout = 0.1
).to(device)

# CCT Vision Transformer
cct = CCT(
    img_size=(128, 128),
    embedding_dim=384,
    n_conv_layers=13,
    kernel_size=3,
    stride=1,
    padding=2,
    pooling_kernel_size=2,
    pooling_stride=2,
    pooling_padding=0,
    num_layers=14,
    num_heads=6,
    mlp_ratio=5.,
    num_classes=10,
    positional_embedding=’learnable’
).to(device)
```

Two models are instantiated (`v & cct`) with specific configurations such as image size, number of channels, patch size, number of classes, depth, number of heads, etc. The model is moved to the specified device (e.g., GPU if available) using `.to(device)`.

D. Loss Function and Optimizer

Cross-entropy loss (`nn.CrossEntropyLoss()`) is chosen as the loss function for training. The Adam optimizer (`optim.Adam`) is utilized for optimizing the model parameters with a learning rate of 0.0001.

E. Training Loop

The training loop runs for a specified number of epochs (`num_epochs`). Within each epoch, the model is set to training mode (`v.train()`), and batches of data are processed. Gradients are zeroed (`optimizer.zero_grad()`), forward pass is computed (`outputs = v(images)`), and loss is calculated (`criterion(outputs, labels)`). Backpropagation is performed (`total_loss.backward()`), and model parameters are updated (`optimizer.step()`). Training loss and accuracy are computed and printed for each epoch.

This methodology outlines the complete pipeline from dataset loading to model training, including data preprocessing, dataset splitting, model setup, loss function definition, optimizer selection, and training loop execution. It serves as a comprehensive guide for training a ViT model for land classification using the EuroSAT dataset.

V. Results and Discussion

1) **Epochs**: Indicates the number of training epochs completed during the training process.

2) **Training Loss**: Represents the average loss (error) computed over the training dataset during the training process. A lower training loss indicates better convergence of the model during training.

3) **Training Accuracy**: Denotes the percentage of correctly classified instances in the training dataset. A higher training accuracy suggests that the model is effectively learning from the training data.

4) **Validation Accuracy**: Represents the percentage of correctly classified instances in the validation dataset. It provides an indication of how well the model generalizes to unseen data

5) **Validation Loss**: Indicates the average loss computed over the validation dataset during model evaluation. A lower validation loss suggests that the model is performing well on unseen data.

6) **Test Loss**: Similar to validation loss, test loss represents the average loss computed over the test dataset during model evaluation. It provides insight into the model's performance on unseen data.

7) **Test Accuracy**: Denotes the percentage of correctly classified instances in the test dataset. It indicates the model's overall performance on unseen data.

8) **Correct Predictions (test)**: Represents the total number of correctly classified instances in the test dataset. It provides a more granular view of the model's performance on individual instances.

9) **Total Predictions**: Indicates the total number of instances in the test dataset. It serves as the denominator for calculating accuracy metrics.

10) **Class Labels**: Classes when extracted from the dataset were given numeric values in order to classify images more conveniently. They are as follows:

```
[0] - AnnualCrop
[1] - Forest
[2] - HerbaceousVegetation
[3] - Highway
[4] - Industrial
[5] - Pasture
[6] - PermanentCrop
[7] - Residential
[8] - River
[9] - SeaLake
```

Comparing the results from the Base ViT and CCT (ViT) models, it is evident that the Base ViT model substantially outperforms the CCT (ViT) model in phrases of training and validation metrics. The Base ViT model performed better with a training accuracy of 90.35% and validation accuracy of 79.15% compared to the CCT (ViT) model, which best achieved training accuracy of 11.16% and validation accuracy of 10.04%. Additionally, the Base ViT model had lower training loss (0.2715) and validation loss (0.7923) compared to the CCT (ViT) model, indicating higher convergence and generalization. This discrepancy between training/validation and test performance could indicate potential overfitting in the Base ViT model or underfitting in the CCT (ViT) model, warranting further investigation into model architecture and training procedures.

TABLE I: Base ViT Result Metrics

Metric	Value
Epochs	15
Training Loss	0.2715
Validation Loss	0.7923
Test Loss	0.6769
Training Accuracy	90.35%
Validation Accuracy	79.15%
Test Accuracy	78%
Correct Predictions(test)	2106
Total Predictions(test)	2700

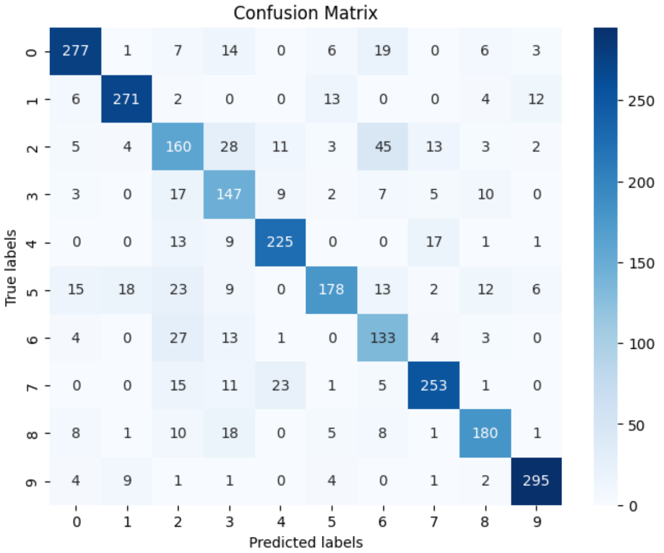
TABLE II: CCT (ViT) Result Metrics

Metric	Value
Epochs	15
Training Loss	4.2636
Validation Loss	3.2567
Test Loss	0.6769
Training Accuracy	11.16%
Validation Accuracy	10.04%
Test Accuracy	11.81%
Correct Predictions(test)	319
Total Predictions(test)	2700

VI. Conclusion

The implementation of the Vision Transformers represents a sophisticated solution characterized by its user-centric design and robust functionality. By adeptly detecting and interpreting Land Classes, the system seamlessly predicts the Land Use, facilitating efficient control over a diverse range of Satellite Image applications. This refined process underscores the system's commitment to optimizing user experience and operational efficacy.

Looking towards the future, there exists considerable potential for further refinement and customization. Specifically, the integration of user-specific datasets holds promise for enhancing the system's adaptability and personalization. By allowing users to train the system with their own datasets, a more tailored and intuitive interaction experience can be achieved, catering to individual preferences and requirements with precision and sophistication. Such advancements not only elevate the system's utility but also reinforce its position at the forefront of smart home automation technologies.



	0	1	2	3	4	5	6	7	8	9
0	277	1	7	14	0	6	19	0	6	3
1	6	271	2	0	0	13	0	0	4	12
2	5	4	160	28	11	3	45	13	3	2
3	3	0	17	147	9	2	7	5	10	0
4	0	0	13	9	225	0	0	17	1	1
5	15	18	23	9	0	178	13	2	12	6
6	4	0	27	13	1	0	133	4	3	0
7	0	0	15	11	23	1	5	253	1	0
8	8	1	10	18	0	5	8	1	180	1
9	4	9	1	1	0	4	0	1	2	295

Fig. 4: Confusion Matrix: Predicted VS Ground Truth

References

[1] O. Adedeji, P. Owoade, O. Ajayi, and O. Arowolo. Image augmentation for satellite images, 2022.

[2] M. Aksoy, B. Sirmacek, and C. Ünsalan. Land classification in satellite images by injecting traditional features to cnn models. *Remote Sensing Letters*, 14(2):157–167, Jan. 2023.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[4] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi. Escaping the big data paradigm with compact transformers, 2022.

[5] S. Kunwar and J. Ferdush. Mapping of land use and land cover (lulc) using eurosat and transfer learning, 2023.

[6] F. T. Lisa, M. Z. Hossain, S. N. Mou, S. Ivan, and M. H. Kabir. Land cover and land use detection using semi-supervised learning, 2022.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers distillation through attention. 2021.

[8] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.