Implementing Convolutional Neural Networks and Vision Transformers for Satellite Image Processing

Matei-Octavian Țurcan, Olesea Caftanatov

Abstract

This paper explores the application of Vision Transformers (ViT) for satellite imagery classification, inspired by the widespread use of transformers in language processing and related computer vision research. The study involved training a ViT model on the EuroSAT dataset and comparing its performance with ResNet50. Additionally, a hybrid model architecture integrating CNNs with ViTs was proposed to leverage the strengths of both approaches.

Keywords: Visual transformers, convolutional neural networks, satellite image, datasets, image recognition.

1 Introduction

In recent years, deep learning has revolutionized the field of computer vision, with convolutional neural networks (CNNs) playing a pivotal role in this success [7]. CNNs have excelled in visual tasks largely due to the powerful inductive bias of spatial equivariance encoded by their convolutional layers. This capability has enabled CNNs to learn general-purpose visual representations that can be easily transferred across various tasks, resulting in consistently strong performance.

However, a significant breakthrough in the field has emerged with the introduction of Vision Transformers (ViTs) [1]. These models, which are adapted from Transformer neural networks originally

©2024 by Matei-Octavian Țurcan, Olesea Caftanatov

designed for language processing, have demonstrated that they can achieve performance on par with, or even surpass, CNNs in large-scale image classification tasks. Unlike CNNs, which aggregate spatial information through convolution, ViTs use self-attention mechanisms to capture information across the entire image, operating in a manner more akin to how Transformers process language data.

Despite their success on large-scale datasets, ViTs often struggle with performance when trained on smaller datasets. In contrast, CNNs like ResNet-18 can achieve high accuracy on small datasets and generally improve as the number of parameters and layers increases.

To address these challenges, we propose a hybrid architecture that combines the strengths of both Convolutional Neural Networks (CNNs) and ViTs. CNNs are well-suited for extracting local features from small and imbalanced datasets, while ViTs excel at capturing global context and complex patterns. By integrating these two approaches, we can achieve more robust and accurate classification across a broader range of satellite image datasets, effectively leveraging the strengths of both models to overcome the limitations inherent in small satellite datasets.

2 The EuroSAT dataset

The dataset used in this experiment is EuroSAT, containing 10 image classes provided by the Sentinel-2 satellite mission. The set is freely available for use under an open licence, as are the images supplied by the satellite. The Sentinel-2 satellite system provides approximately 1.6 terabytes of compressed imagery daily. The data is collected from several countries across Europe to different degrees, dominated by western European countries. The classes are as follows: Forest, River, Highway, Annual Crop, Seas or Lakes, Herbaceous Vegetation, Industrial, Residential, Permanent Crop, and Pasture. This breadth of image classes provides ample opportunity to create specific applications for land cover and use, particularly in agricultural or wild environments with seven out of three classes being reserved for such land uses.

This set contains a relatively large number of images inside each

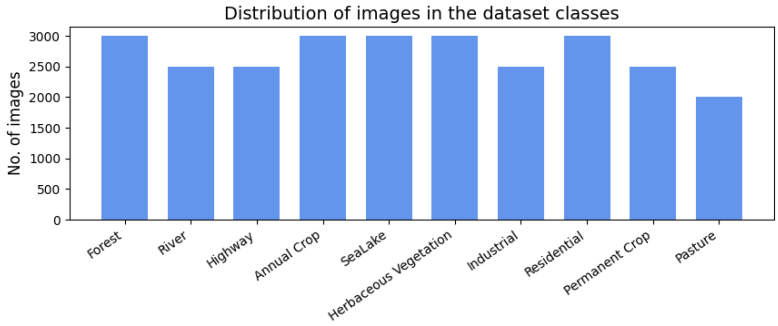


Figure 1. Number of images per each class inside the dataset

class, ranging from 2000 to 3000 as seen in Figure 1, containing a total of 27000 images with a 64 x 64 pixel size. The images were selected based on their clear visibility at a resolution of 10 meters per pixel and were sufficiently covered by the European Urban Atlas. This data set is available in a multi-spectral version, containing 13 spectral channels, and a regular 3 channel RGB set. While the additional spectral channels can be helpful to combat atmospheric effects and provide additional information undetectable by humans to a classification model, this greatly increases the input shape of the images, and consequently, increasing the total training time. As such, the 3 channel set was used in this paper.

The EuroSAT dataset is valuable because it includes various types of land cover, which are essential for urban planning, environmental monitoring, and resource management. By examining changes in urban areas over time, the dataset aids in tracking urban sprawl, identifying urbanization trends, and promoting sustainable city planning. Additionally, the dataset helps monitor crop health, estimate yields, and manage agricultural resources more effectively by analyzing different land cover types and their temporal changes.

3 Vision Transformers

Vision transformers (ViT) are somewhat infrequently used in the field of computer vision for classification tasks, but provide an interesting proposal in terms of handling large datasets with ease. Furthermore, the self-attention trait can prove to be quite useful for global feature detection, as seen in numerous natural language processing applications. The visual equivalent of tokens in the vision transformer architecture is echoed by patches, formed by splitting the desired image into many subsections.

3.1 The Vision Transformer Architecture

The way vision transformers are designed closely resembles how natural language processing (NLP) transformers are created, but slightly modified to accommodate the usage of images in its input.

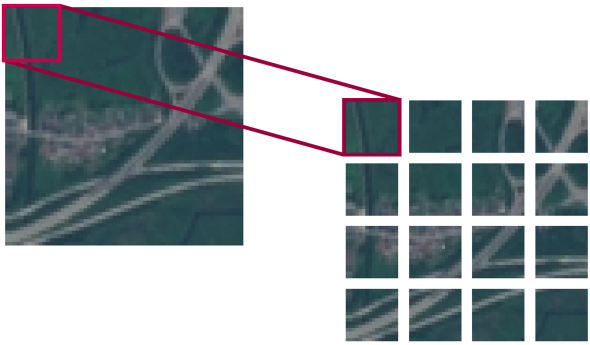


Figure 2. Creation of patches from an input image

Following the architecture of Dosovitskiy et al. [1], the input image is patched into subsections, as seen in Figure 2, which are the visual equivalent of tokens in NLP transformers. The patch size is also defined as a hyper parameter able to be changed in order to determine how broad or narrow the focus should be on the model's local attention.

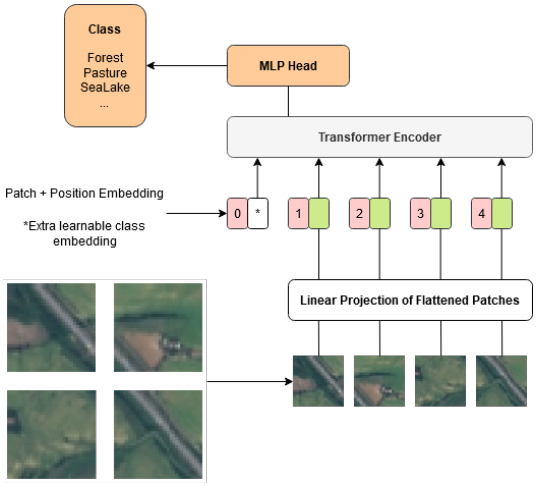


Figure 3. The vision transformer structure

The patches are flattened and are sent into the encoder along a position embedding to determine the patch's position in the original image, and the class embedding used for classification later on in the Mulit-Layer Perceptron (MLP) head.

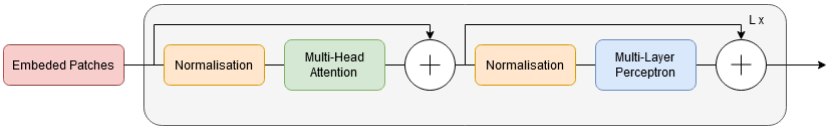


Figure 4. The transformer encoder architecture

Inside the Multi-Head Self-Attention block is the mechanism by which the model is able to focus on the important details found in the image, enabled by learning the relationships between the patches which are fed into the encoder.

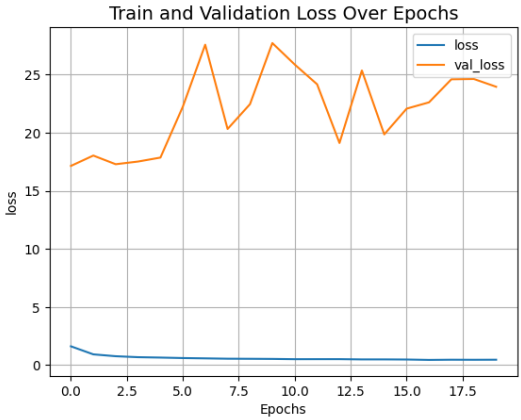


Figure 5. Vision transformer validation loss

3.2 Using ViTs for direct classification

Dosovitskiy et al. [1] have used ImageNet-21k and the JFT-300M datasets to pre-train the ViT model stating a great degree of success. However, both ImageNet-21k and JFT-300M are very large datasets with over 21000 and 375 million general classes respectively compared to EuroSAT. Additionally, both sets have a total number of pictures several orders of magnitude higher than the EuroSAT dataset, with the images consisting of a larger resolution too. All these characteristics result in a much more lessened performance proposal in computer vision for ViTs in the context of satellite imagery.

4 Testing ViTs on EuroSAT

The previously mentioned lack of performance due to the small scale of the dataset, is reflected by its performance across multiple test runs of the model. Initially, the model was very unstable and erratic in output; however, slightly changing the model has resulted in a low 11% validation accuracy.

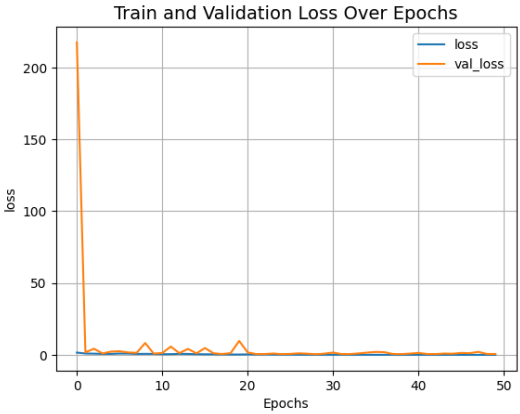


Figure 6. ResNet50 validation loss

The graph in Figure 5 indicates that the ViT has issues with a lot of overfitting, and this is further proof of the data-hungry characteristics of the model along the other factors discussed by Zhu et al. [6].

Training the model from scratch to only classify image data proves to be complicated and unreliable in performance, but this could be mitigated by creating an even larger dataset. This shows that there is a compromise to be had regarding the retrieval of satellite data and ViT performance, as attaining such detail from satellite imagery can become cost-prohibitive with greater spatial resolution. To contrast this, the satellite used here provides free access to its data, making it much more accessible.

While the ViT trains slightly quicker than ResNet50, the overall testing accuracy result is not up to the same level as the CNN mentioned. When training ResNet50 by itself, it achieved a validation accuracy of 86% in 50 epochs, with the validation loss on the final epoch staying at 0.62, and the overall low loss of the final epoch is also reflected in Figure 6. This is a good result that could be used standalone, but the model started to show diminishing returns over the last epochs.

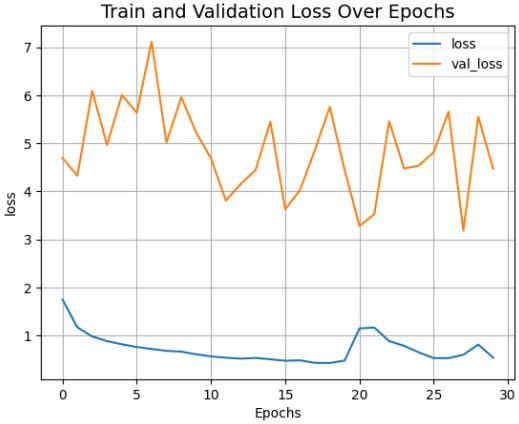


Figure 7. Hybrid model validation loss

4.1 The hybrid model approach

With these results, a hybrid approach to integrating ViTs in image classification tasks alongside CNNs can be a show of strength for both types of models, especially considering the previous limitations of using a ViT by itself. This approach was proposed by Dosovitskiy et al. [1], however, as mentioned by Ngo et al. [5], there are several ways to both training and implementing ViTs with CNNs. For this use case, the ViT is used to replace the final layers of the model ResNet50 to handle the image classification. This approach ensures minimal conflicts between input shapes for the layers used.

Similar to the individual ViT, the overall performance of the hybrid model is very poor. The validation loss shown in Figure 7, and the total accuracy of 12% shows that even a feature extractor like ResNet50 cannot improve the classification aspects of a vision transformer on low-level applications. Furthermore it proves that overfitting still occurs even when coupled with the efficiency of CNNs on small datasets.

5 Conclusion & Future work

Vision transformers are not particularly effective for classifying satellite images when used alone or in combination with convolutional neural networks on small datasets. Although combining vision transformers with convolutional neural networks generally mitigate some of the individual weaknesses of each architecture on larger sets, the overall performance remains ineffective for small datasets. While this raises concerns about the viability of vision transformers in such scenarios, it also presents an opportunity to explore their integration as a complementary component embedded within convolutional neural networks and to investigate further optimization techniques tailored for small datasets.

Acknowledgments. SIBIA - 011301, Information systems based on Artificial Intelligence has supported part of the research for this paper.

References

[1] 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*. In Proceedings of International Conference on Learning Representations, 03-07 May, 2021, Austria. pp. 1–22, `https://doi.org/10.48550/arXiv.2010.11929`.

[2] P. Helber, B. Bischke, A. Dengel, and D. Borth. *EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification*, 2017, pp. 1–9. [Online] `https://doi.org/10.48550/arXiv.1709.00029`. [Accessed 07.08.2024] Dataset download: `https://zenodo.org/records/7711810#.ZAm3k-zMKEA`.

[3] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*, 2015, pp. 1–12. [Online] https://doi.org/10.48550/arXiv.1512.03385. [Accessed 08.08.2024].

[4] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. *ImageNet-21K Pretraining for the Masses*, 2021, pp. 1–3. [Online] https://doi.org/10.48550/arXiv.2104.10972. [Accessed 10.08.2024].

[5] B. H. Ngo, N. Do-Tran, T. Nguyen, H. Jeon, and T. J. Choi. *Learning CNN on ViT: A Hybrid Model to Explicitly Class-specific Boundaries for Domain Adaptation*, 2024, pp. 1–15. [Online] https://doi.org/10.48550/arXiv.2403.18360 [Accessed 10.08.2024].

[6] H. Zhu, B. Chen, and C. Yang. *Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective*, 2023, [Online] https://doi.org/10.48550/arXiv.2302.03751 [Accessed 15.08.2024].

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.

Matei-Octavian Țurcan[1], Olesea Caftanatov[2]

[1] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University
E–mail: matei.turcan@math.md
ORCID: https://orcid.org/0009-0004-2232-8186

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University
E–mail: olesea.caftanatov@math.md
ORCID: https://orcid.org/0000-0003-1482-9701