**University of Science and Technology of Hanoi**

The Vision Transformer for remote sensing image classification

Group members:

Nguyễn Thành Nam - 22BI13045

Trần Trọng Nghĩa - 22BI13332

Trần Đức Dương - 22BI13117

Nguyễn Tuấn Dũng - 22BI13107

Trương Đại An - 22BI13008

Vũ Hùng Anh - 22BI13045

**September 2024**

**Introduction**

*Problem Statement*

The classification of satellite images into various categories plays a crucial role in multiple fields, including environmental monitoring, urban planning, and disaster management. Accurate classification helps in making informed decisions by providing precise land use and cover information. This study investigates the use of Vision Transformers (ViT) for satellite image classification, leveraging their advanced architecture to achieve superior performance compared to traditional methods.

*Background*

Traditional methods for satellite image classification often rely on Convolutional Neural Networks (CNNs), which, while effective, have limitations in capturing long-range dependencies within images. Vision Transformers, initially designed for natural language processing tasks, offer a novel approach by treating image patches as sequences, enabling the model to recognize complex spatial patterns.

*Objectives*

The primary objective of this study is to implement and evaluate the performance of Vision Transformers in classifying satellite images. We aim to demonstrate their effectiveness in handling diverse land use and cover types, and explore potential improvements over existing methods.

**Dataset**

The EuroSAT dataset, used in this study, consists of RGB images representing various land use and land cover types across Europe. It provides a comprehensive set of scenarios, including residential, industrial, and agricultural areas, enabling thorough testing of classification models. The dataset includes 27,000 labeled and geo-referenced images, each covering multiple spectral bands.

*Data Characteristics*

The dataset's diversity in classes allows for a robust evaluation of model performance. Each image is associated with one of ten classes, providing a rich variety of features for the model to learn from. The inclusion of geo-referenced data adds an additional layer of complexity and utility, facilitating applications in geographical information systems (GIS).

**Methodology**

*Data Preprocessing*

Effective data preprocessing is crucial for enhancing model performance. The following techniques were employed:

* Data Augmentation: To increase data variability and improve model robustness, augmentation techniques such as rotation, flipping, and scaling were applied. These methods help simulate different observational conditions, reducing overfitting by exposing the model to a broader range of scenarios during training.
* Normalization: Standardizing pixel values is essential for improving model convergence and stability. By ensuring that input data falls within a consistent range, normalization helps in optimizing the learning process, allowing the model to focus on relevant patterns rather than data discrepancies.

*Model Architecture*

The Vision Transformer (ViT) architecture represents a shift from traditional CNN-based models. It utilizes a transformer framework, originally developed for NLP tasks, to handle image classification. The key aspects include:

* Patch Embedding: Images are divided into small patches, each treated as a token. This process transforms the image into a sequence of embeddings, similar to how words are processed in NLP models.
* Patch Embedding Formula:

E = Linear(x) + Position\_Encoding

* Transformer Encoder: The model processes these sequences through multiple transformer encoder layers. Each layer consists of a multi-head self-attention mechanism and feed-forward neural networks, enabling the model to capture intricate patterns and dependencies across the image.

*Training Process*

The training process involves careful tuning of hyperparameters to optimize model performance:

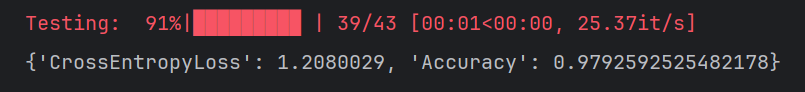
* Hyperparameter: The learning rate and batch size are tuned to optimize performance. The Adam optimizer is used for efficient gradient descent, crucial during model training.
* Training Duration: The model undergoes training over multiple epochs to ensure convergence. This iterative process allows the model to incrementally improve its accuracy and generalization capability.
* Validation: Regular evaluations on a separate validation set are conducted to monitor overfitting. Early stopping and learning rate adjustments are employed based on validation performance to enhance model reliability.

**Results**

The effectiveness of Vision Transformers is demonstrated through their performance on the test set. The model achieved a high classification accuracy, highlighting its capability to handle complex satellite imagery tasks.

*Accuracy and Loss*

* Accuracy: The model's classification accuracy reflects its ability to correctly identify land use and cover types. Detailed results are provided in the accompanying image.
* Loss Function Formula: Loss =
* Loss: Throughout training, loss metrics were monitored to ensure proper model learning.

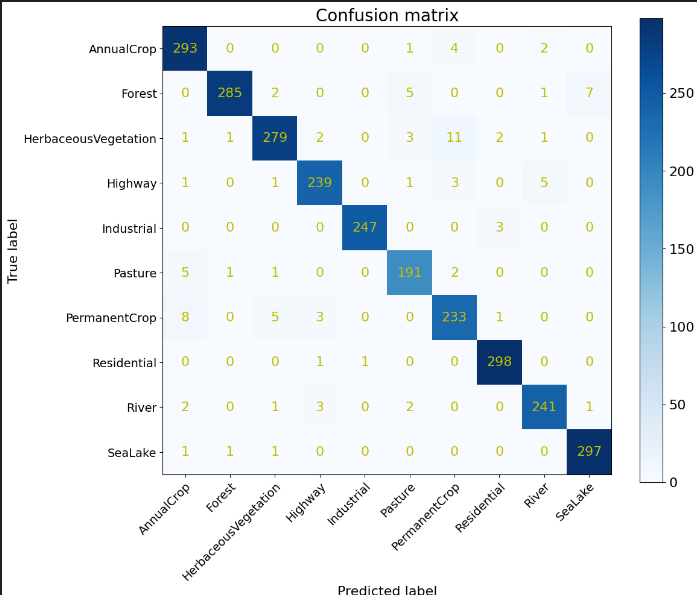
****

****

*Evaluation Metric*

To comprehensively evaluate the model's performance, several metrics were used:

* Confusion Matrix: This tool provides a visual representation of the model's performance across different classes, highlighting areas of strength and potential improvement.



* Precision, Recall, F1-Score: These metrics offer insights into the model's precision and sensitivity, providing a nuanced understanding of its classification capabilities.

**Discussion**

*Comparative Analysis*

Comparing the ViT model with traditional CNN approaches reveals significant improvements in handling long-range dependencies and complex patterns. The transformer-based architecture offers flexibility in adjusting to diverse image characteristics, making it suitable for various remote sensing applications.

*Practical Implications*

The successful implementation of Vision Transformers in satellite image classification opens new possibilities for real-world applications. Enhanced accuracy in land use and cover classification can aid in timely decision-making for urban planners, environmentalists, and disaster response teams.

*Limitations and Challenges*

While ViTs offer notable advantages, they also present challenges such as increased computational requirements and sensitivity to hyperparameter tuning. Future research should focus on optimizing these aspects to make the models more accessible and efficient.

**Conclusion**

* Summary: Vision Transformers have demonstrated remarkable effectiveness in classifying satellite images, offering significant improvements over traditional methods. Their ability to capture complex spatial patterns makes them a valuable tool in remote sensing.
* Future Work: Explore further improvements with larger datasets and more complex architectures.

**Example:**

**References:**

[EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification](https://zenodo.org/records/7711810?fbclid=IwZXh0bgNhZW0CMTAAAR3AM7cY_VfIwwIxIfEyJyH2ck7zLlahOdw7cTOm0-GhLT9ibAaGzRrDpTM_aem_e6Y_4JxKXU8WaIVqK-aQPg#.ZAm3k-zMKEA)