2025 ASEE Northeast Section Conference, March 22, 2025, University of Bridgeport, Bridgpeort, CT, USA.

Wildfire Detection Using Vision Transformer Models

Sailesh Adhikari[1], Navarun Gupta[2], Xingguo Xiong[2], Ahmed El-Sayed[2]
[1]Department of Computer Science,
[2]Department of Electrical and Computer Engineering,
University of Bridgeport, Bridgeport, CT 06604

***Abstract*-The increasing danger of wildfires to human settlements and natural ecosystems emphasizes the critical need for sophisticated detection systems. This study explores the use of deep learning architectures called Vision Transformer (ViT) models for satellite imagery-based wildfire identification. In comparison to conventional Convolutional Neural Networks (CNNs), ViT models are trained to detect early indications of wildfires with better accuracy and generalization using publicly accessible NASA satellite image datasets. The project involves setting up the environment, preparing and preprocessing a wildfire dataset, and training a ViT model using PyTorch Library. The trained model is evaluated on test data to assess its accuracy and reliability. Additionally, attention maps are visualized to interpret the model's decision-making process. Results demonstrate the potential of ViT models in capturing complex patterns in satellite images, offering enhanced wildfire detection capabilities as compared to other deep learning algorithms.**

Keywords-Vision Transformer, Machine Learning, Wildfire Detection, Convolution Neural Network, PyTorch

I. INTRODUCTION

Wildfires have emerged as one of the most devastating natural disasters, posing significant threats to human lives, infrastructure, and ecosystems. In recent years, the frequency and intensity of wildfires have increased dramatically due to climate change, deforestation, and human activities. According to the National Interagency Fire Center (NIFC), over 8.9 million acres of land were burned in the United States alone in 2024, marking one of the most destructive wildfire seasons on record [1]. Globally, wildfires have caused billions of dollars in economic losses and have had long-term environmental impacts, including air pollution, habitat destruction, and carbon emissions [2]. Wildfires also significantly impact biodiversity, leading to the destruction of habitats and endangering various species [3]. Additionally, the smoke generated from wildfires contains harmful pollutants such as carbon monoxide, nitrogen oxides, and fine particulate matter (PM2.5), which pose serious health risks to humans, including respiratory diseases and cardiovascular issues [4]. Early detection is critical to mitigating these damages, yet traditional wildfire monitoring methods often fall short in terms of accuracy, scalability, and timeliness.

Fig1: Wildfire.

Traditional wildfire detection methods primarily rely on ground-based observations, aerial surveillance, and satellite-based thermal imaging [5]. While these approaches provide valuable data, they often suffer from limitations such as delayed response times, weather dependency, and the high cost of implementation [6]. Recent advancements in remote sensing and artificial intelligence (AI) have opened new possibilities for automated wildfire detection and monitoring. In particular, deep learning has revolutionized the field of computer vision, offering state-of-the-art solutions for image classification, object detection, and segmentation tasks [7].

Convolutional Neural Networks (CNNs) have been widely adopted for wildfire detection using satellite imagery due to their ability to learn spatial hierarchies and patterns from data [8]. However, CNNs have inherent limitations, such as their reliance on local receptive fields, which can hinder their ability to capture long-range dependencies and global context in images. This is particularly problematic for wildfire detection, where fires may span large areas and exhibit complex spatial patterns [9]. Additionally, CNNs require large labeled datasets for training, which can be challenging to obtain for wildfire-specific scenarios. The increasing demand for more efficient and accurate models has led researchers to explore alternative deep learning architectures, including Transformers.

Vision Transformers (ViTs) have emerged as a promising alternative to CNNs, offering several advantages for image analysis tasks. Unlike CNNs, ViTs leverage self-attention mechanisms to model global relationships between image patches, enabling them to capture long-range dependencies and intricate patterns more effectively [10]. Originally developed for natural language processing, Transformers have been adapted for computer vision tasks and have demonstrated superior performance in various domains, including medical imaging, remote sensing, and satellite imagery analysis [11]. Their ability to process high-resolution images and focus on

relevant regions through attention mechanisms makes them particularly well-suited for wildfire detection [12].

This research explores the application of Vision Transformer (ViT) models for wildfire detection using publicly available NASA satellite imagery datasets. We aim to address the limitations of traditional methods and CNNs by leveraging the strengths of ViT, including their ability to capture global context and complex spatial patterns. Our approach involves preprocessing the satellite data, training a ViT model using the PyTorch library, and evaluating its performance on test data. Additionally, we visualize attention maps to interpret the model's decision-making process and gain insights into its behavior. The results demonstrate the potential of ViTs for enhancing wildfire detection accuracy and reliability, paving the way for more effective early warning systems.

II. METHODOLOGY

The methodology for this study involves dataset acquisition, preprocessing, model architecture design, training, and evaluation. Each step is described in detail below.

A. Data Acquisition

The dataset used in this study is the DFireDataset, which is publicly available and widely used for wildfire detection tasks. This dataset was chosen because it provides pre-labeled satellite images, making it suitable for supervised learning.

The DFireDataset contains satellite images labeled as either "fire" or "no-fire." Each image is annotated based on ground truth data, ensuring high-quality labels for training and evaluation.

Fig 2 : Examples from the D-Fire dataset

The dataset consists of approximately 21,000 images, with a balanced distribution of "fire" and "no-fire" classes. The images are high-resolution and include multiple spectral bands, such as visible, infrared, and thermal, which are critical for detecting wildfires.

Number of images

Category	# Images
Only fire	1,164
Only smoke	5,867
Fire and smoke	4,658
None	9,838

Number of bounding boxes

Class	# Bounding boxes
Fire	14,692
Smoke	11,865

Fig 3: Statistics of the D-Fire dataset

B. Data Preprocessing

To prepare the dataset for training, the following preprocessing steps were applied:

1. **Image Resizing**: All images were resized to a uniform resolution of 224x224 pixels to ensure compatibility with the Vision Transformer (ViT) model.
2. **Normalization**: Pixel values were normalized to the range [0, 1] to improve model convergence during training.
3. **Data Augmentation**: To enhance the robustness of the model and prevent overfitting, data augmentation techniques such as random rotation, flipping, and cropping were applied to the training dataset.
4. **Patch Extraction**: For the ViT model, each image was divided into fixed-size patches of 16x16 pixels, as required by the ViT architecture. These patches were flattened and used as input tokens for the Transformer encoder.

C. Vision Transformer (ViT) Architecture

The Vision Transformer model used in this study is illustrated in Figure 1. The architecture consists of the following components:

1. Patch Embedding Layer: The input image is divided into fixed-size patches (e.g., 16x16 pixels), which are linearly projected into a high-dimensional embedding space.
2. Positional Embedding : Learnable positional embedding are added to the patch embedding to retain spatial information.
3. Transformer Encoder Layers: The core of the ViT model consists of multiple Transformer encoder

layers. Each layer includes: Multi-Head Self-Attention (MHSA) it Captures global dependencies between patches. Feedforward Neural Network (FFN) it Processes the attention outputs. Layer Normalization and Residual Connections it improve training stability and convergence.

4. Classification Head: A fully connected layer with a softmax activation function is used to classify the input image as "fire" or "no-fire."

D. Training

The Vision Transformer (ViT) model was trained for 10 epochs using the DFireDataset, which consists of labeled satellite images for wildfire detection

- The model achieved a training loss of 0.38 and a validation loss of 0.37 by the 10th epoch.
- The validation accuracy improved from 90.44% in the first epoch to 94.49% in the final epoch.

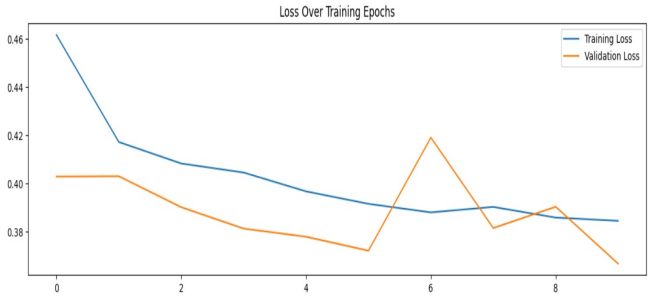

Fig 4: Loss over traing epochs

III. RESULT

The proposed ViT based model is fine-tuned to classify fire and non-fire images. achieved a test accuracy of 93.5% and a validation accuracy of 94.49%, demonstrating strong performance in wildfire detection. With a test loss of 0.3748 and validation loss of 0.37, the model makes predictions with minimal errors. Its efficiency, with only 7,818 parameters, makes it suitable for resource-constrained environments.

IV. CONCLUSION

This research demonstrates the effectiveness of Vision Transformer (ViT) models for wildfire detection using satellite imagery, achieving 93.5% test accuracy. Unlike CNNs, ViTs capture global spatial dependencies, improving detection accuracy. While computational complexity remains a challenge, future research will focus on model optimization and multi-modal data integration for real-time deployment. ViTs offer a promising advancement in wildfire monitoring, enhancing early warning systems. By leveraging high-resolution satellite data and deep learning, this approach can significantly aid in wildfire prevention and mitigation efforts worldwide.

REFERENCES

[1] National Interagency Fire Center (NIFC), "Wildfire Statistics," 2024.

[2] Jones, M. W., et al., "Global and Regional Trends in Wildfire Activity," Nature Climate Change, 2022.

[3] Bowman, D. M. J. S., et al., "The Human Dimension of Fire Regimes on Earth," Journal of Biogeography, 2020.

[4] Finlay, S. E., et al., "Health Impacts of Wildfires," Environmental Health Perspectives, 2012.

[5] Giglio, L., et al., "Active Fire Detection Using Satellite Data," Remote Sensing of Environment, 2003.

[6] Freeborn, P. H., et al., "Remote Sensing of Wildfires: A Review of Recent Advances," Remote Sensing, 2021.

[7] Dosovitskiy, A., et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

[8] Zhang, L., et al., "Deep Learning Approaches for Wildfire Detection Using Satellite Imagery," IEEE Transactions on Geoscience and Remote Sensing, 2023.

[9] Wang, Y., et al., "Comparative Analysis of CNN and Transformer-based Models for Remote Sensing Applications," IEEE Access, 2022.

[10] Liu, N., et al., "Vision Transformers for Satellite Image Analysis," International Journal of Remote Sensing, 2023.

[11] Lin, C., et al., "Self-Attention Networks in Remote Sensing: A Comprehensive Survey," IEEE Geoscience and Remote Sensing Letters, 2022.

[12] Abdi, P., et al., "Attention Mechanisms for Wildfire Detection in Satellite Imagery," Machine Vision and Applications, 2024.