Aerial Wildfire Detection using Image Classification

Firat Taxpulat
Stanford University
735 Campus Drive, Stanford, CA
firattax@stanford.edu

Abstract

An increasingly prominent issue in the world are wildfires. Wildfire detection can help save properties, homes, and ultimately lives. In this paper, convolutional neural networks and transformers are used to detect wildfires using aerial images captured by drones and other unmanned aircraft. The two architectures used achieve accuracy scores of 80% and 89%

1. Introduction

Wildfires have increasingly become a significant environmental and economic threat worldwide, causing devastating impacts on ecosystems, human lives, and property. For instance, several catastrophic wildfire events recently occurred in Southern California, particularly in my hometown region of Lancaster, Palmdale, and the rest of the Antelope Valley area. In September 2023, the Fairmont Fire near Palmdale and Lancaster rapidly spread which led to extensive firefighting efforts and emergency evacuations for my family and other Antelope Valley residents. The event highlighted the region's growing vulnerability to wildfires and the need for advanced detection technologies to aid in early intervention for my hometown and other regions prone to wildfires around the world.

NASA and other space agencies have developed satellite systems that use thermal imaging to detect wildfires from space. Unfortunately, satellite systems are unable to accurately monitor small brush fires that have the potential to spread and grow into larger wildfires. This along with the high maintenance costs of satellite systems lead satellite systems to be an unscalable solution for wildfire detection.

Aerial drones provide a low-cost, effective solution to wildfire detection. Equipped with cameras, aerial drones can monitor large areas quickly, detect both small and large wildfires, and provide real-time data to emergency responders.

This report delves into the application of convolutional neural networks (CNNs) and transformers for detecting wildfires from aerial images. We will use a CNN model derived from a combination of transfer learning and additional training techniques as a baseline model for wildfire detection. Then, we will compare the results of the CNN model and a transformer model built using only transfer learning. The results of this report will convey the feasibility of integrating such computer vision models to aerial drones to solve the problem of wildfire detection.

1.1. Problem Statement

The problem involves predicting whether a wildfire is present in images taken by an aerial drone. This is a binary classification problem, and the previously described baseline model and experiment model predict the two classes: *Fire* and *No Fire*. When making a prediction, the models output classification scores for both classes for a given input image that consists of either burning forests or non-burning forests. The baseline and experiment models are evaluated using classification accuracy and F1 score.

2. Related Works

2.1. Utilizing Convolutional Neural Networks for Detecting Wildfires

Wildfire detection using CNN image classification has made advances in the past. In *Deep Learning Approaches for Wildland Fires Remote Sensing: Classification, Detection, and Segmentation*[1], Ghali et al. uses segmentation architectures to classify wildfire images. Of the different segmentation architectures discussed in the study, the most noteworthy is U-Net. Initially designed for segmentation on biomedical images, U-Net incorporates an encoder-decoder structure that features taking the outputs of some layers in the neural network, skipping subsequent layers, and using these outputs as the inputs of later layers. In the Ghali et al. paper, U-Net is applied on the *FLAME* dataset that contains wildfire images taken by drones in Arizona forests. From U-Net's effective segmenting of complex shapes and fine details for classifying biomedical images, Ghali et al. uses U-Net to parse notable wildfire features possibly present in an image.

Another study on wildfire detection is *Attention based CNN model for fire detection and localization in real-world images* [2] by Majid et al. that utilizes an attention-based CNN architecture. In the Majid et al. paper, the datasets used are not explicitly named as the authors state, “a composite dataset was created by collecting images collected from the data of well-known public datasets used in recent works on this problem”.

The attention-based CNN is designed starting with the EfficientNet-B0 neural network, and an attention mechanism is integrated into the CNN. This attention mechanism tunes the model to focus on the regions of an image most likely to contain a wildfire based on the presence of wildfire features in those regions, like smoke, sparks, and burnt vegetation.

A related study is *EmergencyNet: Efficient Aerial Image Classification for Drone-Based Emergency Monitoring Using Atrous Convolutional Feature Fusion*[3] by Kyrkou et al. where they propose an approach that efficiently classifies aerial images for monitoring emergency situations using drones. This approach is particularly advantageous due to its quick and efficient classification allowing drones without significant on-board computing power to perform wildfire classification using aerial images. In the Kyrkou etl al. paper, the approach is applied on a dataset of images containing different natural disasters the authors manually compiled from other datasets.

The method discussed in the work is known as ACFF, Atrous Convolutional Feature Fusion. This method involves using atrous convolution layers instead of standard convolution layers in all but the first and last layer of a neural network. The atrous convolution layers capture the same contextual information as the standard convolution layers without significantly increasing the computational load. This is done by spacing out filter parameters through introducing gaps between them that are controlled by a dilation rate. This process enlarges the receptive field of the atrous filters without increasing the parameters of the neural network model, thereby maintaining effective feature extractors of images at a lower computational cost than standard convolutions.

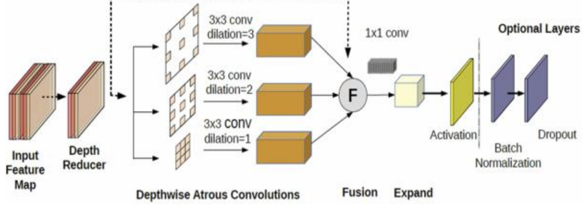


Figure 1: **Basic components of Atrous Fusion Block**

.

Hong et al. employ a ConvNeXt model for wildfire detection detailed in their paper *Wildfire Detection Via Transfer Learning: A Survey*[4]. ConvNeXt is a convolutional neural network architecture that integrates layers that assist with CNN generalizability, like depthwise convolutions and layer normalization. It simplifies traditional CNN designs while achieving high performance results in image classification and other computer vision tasks.

The baseline CNN model mentioned previously is based on this ConvNeXt model and will later be compared with the results of the CNN models discussed prior.

2.2. Applying Transformers on Aerial Images

In addition to applying convolutional neural networks toward detecting wildfires, recent studies on transformer applications for wildfire detection have been made. The paper *Early fire detection technology based on improved transformers in aircraft cargo compartments* [5] by Ai et al. addresses the issue of high false alarm rates associated with fire detection systems by improving on transformers’ long-term dependency handling capabilities.

The authors created a transformer integrated with a multi-head self-attention mechanism to capture long-range dependencies. These transformers are enhanced with contextual embeddings enabling the model to interpret complex environmental features in an image, i.e. haze or smoke, orange leaf trees or burning trees etc. This improvement helps with distinguishing actual fires from false positives. Additionally, the model uses a thresholding mechanism that adjusts the thresholds for determining a fire being present in an image based on real-time environmental data, like weather, air quality, moisture level etc. This further reduces the rate of false alarms being determined by the authors’ transformer.

In the paper *Transforming Wildfire Detection and Prediction Using New and Underused Data Sources Integrated with Modeling*[6] the authors Coen et al. approach the wildfire detection problem differently than the previously mentioned papers. The study uses dynamic data-driven application systems (DDDAS) and models of weather-fire behavior for wildfire detection. This study utilizes transformers to analyze diverse sources of wildfire data, including satellite images, sensors, and social media reports, to improve the accuracy and timeliness of wildfire detection.

The transformer model uses a multi-head attention mechanism to capture relationships between different types of data. For example, social media reports can provide real-time information about the locations of emerging wildfires while satellite imagery offers a broad overview of areas affected by wildfires. Sensors provide data closely related to fires, like temperature, humidity, and wind conditions. By converting these diverse data sources as embeddings and feeding them to the transformer's multi-head self-attention layers, the paper's transformer model generates a holistic view of possible forest fire occurrences which improves the effectiveness of wildfire detection systems.

Adding to their work on wildfire classification using CNNs discussed in section 2.1, Ghali et al. study the applications of transformers toward detecting wildfires in *Deep Learning and Transformer Approaches for UAV-Based Wildfire Detection and Segmentation*[7]. Their work compares traditional deep convolutional networks, like MobileNetV3, DenseNet, and InceptionV3, with their vision transformers developed for aerial wildfire detection. The model with the strongest performance in their study is TransUnet, a transformer model combined with the U-Net CNN architecture discussed previously in this report. TransUnet integrates a ResNet50 feature extractor into a pre-trained Vision Transformer (ViT). Passing these feature maps into a decoder, the model concatenates the features with the output of ResNet50.

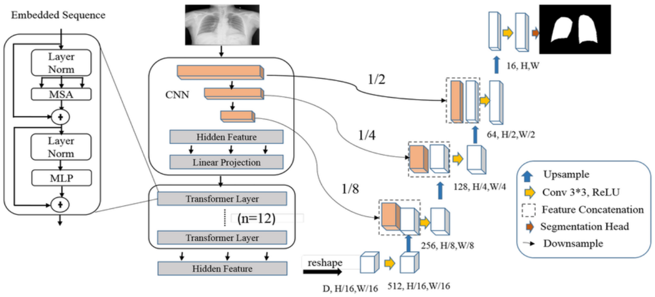


Figure 2: TransUnet Architecture

Building on the foundation laid by their previous paper, Ghali et al. delve deeper into the capabilities of TransUNet for forest fire detection in the study *Wildfire Segmentation Using Deep Vision Transformers*[8]. The study introduced multiple improvements to TransUnet to optimize the model for wildfire detection. One improvement is adding positional encodings dynamic to the different amounts of fire imagery present in an image. Another improvement is including thresholding techniques to enhance TransUNet's sensitivity to subtle signals of fire amidst backgrounds containing noisy elements. Additionally, Ghali et al. employ random cropping, rotation, and brightness adjustments to augment their datasets to improve the TransUNet model's generalization capability.

In *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*[9], Liu et al. uses the Swin Transformer architecture to efficiently analyze images. This architecture divides the input image into non-overlapping windows, thereby enabling the Swin Transformer model to process images using both local and global contexts. The windows are then shifted, ensuring that the model can capture interactions across different parts of the image, which is crucial for detecting wildfires that can vary significantly in size and shape.

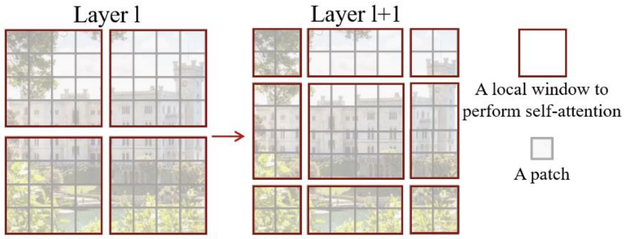


Figure 3: Visualization of shifted window approach in Swin Transformer architecture. The left image (layer l) illustrates a non-shifted window partitioning scheme where self-attention is computed within each window. The right image (layer l + 1) shows a shifted window partitioning where new windows are created. Self-attention computations in the new windows overlap with the previous layer's self-attention computations

Rekavandi et al. discuss the ability to detect small objects in their paper *Transformers in Small Object Detection: A Benchmark and Survey*[10]. Their work involves refining the patch embedding process used in transformer models. Typically, images are divided into 32x32 pixel patches, but Rekavandi et al. note for small object detection, smaller patches (e.g., 16x16 pixels or less) may be used to ensure finer granularity. Each patch is then embedded into a fixed-size vector through a linear projection, thereby preserving detailed information crucial for recognizing features of small objects, like small fires, amidst complex backgrounds.

The experimental transformer model mentioned previously is based on the last two discussed transformer model studies. This experimental transformer model will later be compared with the results of the baseline CNN model as well as the CNN models discussed in prior mentioned papers.

3. Methodology

3.1. Baseline CNN Model

As stated previously, the baseline CNN model is based on the ConvNeXt model detailed in *Wildfire Detection Via Transfer Learning: A Survey*[4]. The authors utilized a ConvNeXt model derived from TensorFlow that is pre-trained on the ImageNet-1K dataset for forest fire detection. The training involved using 15 epochs where a fixed feature extractor method is used in the first 10 epochs and a transfer learning method is used in the last 5 epochs. Wildfire classification was determined by feeding the results of the ConvNeXt model to a softmax layer and using the binary cross-entropy loss function to measure the model's efficacy.

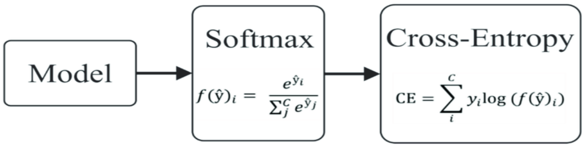

Model
Softmax
$f(\hat{y})_i = \frac{e^{\hat{y}_i}}{\sum_j^C e^{\hat{y}_j}}$
Cross-Entropy
$\mathrm{CE} = \sum_i^C y_i \log(f(\hat{y})_i)$

Figure 4: Visualization of model–softmax–loss architecture. Due to wildfire detection being a binary classification problem, i.e. fire or no fire, the loss function above reduces to the binary cross-entropy loss function

The baseline ConvNeXt model developed using PyTorch follows this transfer learning design by using the combined wildfire dataset to train the last layer of the ConvNeXt model after 10 epochs. This baseline model differs from the ConvNeXt model in the previously mentioned paper as the baseline model will train for 20 epochs instead of 15 epochs Then, the baseline model is applied to the validation set to evaluate its accuracy and F1 scores. The results of the validation set are derived from varying dropout and learning rate to hyperparameter tune the model using randomized search. Finally, the model's effectiveness is determined by its accuracy and F1 scores on the testing set.

3.2. Experimental Transformer Model

The experimental transformer model for wildfire detection is developed for this report by integrating techniques from the previously mentioned papers: *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*[9] and *ViT: Vision Transformer for Small-Size Fire Detection*[10]. By combining the efficient analyzing of images provided by the Swin Transformer architecture with the fine granularity given by using small patches, the experimental transformer model has a robust wildfire detection system.

For our purposes, the experimental transformer model begins as the Swin Transformer pre-trained on the ImageNet-1K dataset. Then, transfer learning is employed to the model to train the last layer of the Swin Transformer using the combined wildfire dataset. Throughout this training, the experimental transformer model employs using smaller patches, like 16x16 pixels, instead of the usual size patches, like 32x32 pixels, of the Swin Transformer.

Similar to the baseline CNN model after being trained with the combined wildfire dataset, the experimental transformer is applied to the validation set to evaluate its accuracy and F1 scores. The results of the validation set are derived from varying dropout and learning rate to hyperparameter tune the model using randomized search. Finally, the model's effectiveness is determined by its accuracy and F1 scores on the testing set.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Figure 5: Accuracy and F1 score equations used for both baseline CNN model and experimental transformer model; TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative

3.3. Ensemble Learning

In this report, ensemble learning is used to enhance the generalizability robustness and accuracy of wildfire detection. The architecture of the baseline ConvNeXt model remains the same as previously discussed with now using three ConvNeXt models trained independently from one another. Each ConvNeXt model continues to be trained by varying the hyperparameters dropout rate and learning rate using random search as indicated previously. Each ConvNeXt model is still pre-trained on the ImageNet-1K dataset and fine-tuned on the wildfire dataset developed for this report.

During inference, the predictions from each of the three ConvNeXt models are combined using a voting scheme. The final prediction is determined by averaging the outputs from the models and selecting the class with the highest average score. By utilizing ensemble learning, the aggregated ConvNeXt model leverages the strengths of multiple models to improve overall classification performance.

4. Dataset

The dataset used for training, validating, and testing is comprised of four different datasets: *FIRE Dataset*[11], *Wildfire Detection Image Data*[12], *FlameVision*[13], and *The wildfire dataset*[14]. Each dataset is composed of *Fire* and *No Fire* images; however, some datasets are organized differently than others. For example, *FlameVision* and *The wildfire dataset* are already organized into training, validation, and testing set images whereas the *Fire Dataset* and *Wildfire Detection Image Data* datasets are not.

Images in these datasets are combined and separated as either *Fire* or *No Fire* images regardless of images from some datasets being exclusively organized as training, validation, or testing set images. Then, the combined dataset is split into a 60:20:20 ratio for training, validation, and testing, respectively. This ratio is used to provide abundant training data for the baseline and experiment models as well as having ample data to evaluate the performance of the trained models.

A variety of wildfire images are provided in the combined dataset. This variety includes images with large smoke plumes, enormous blazing wildfires, small smoldering brushfires, and many more distinct wildfire features. By training the baseline and experimental models on a diverse range of wildfires, the generalizability of the baseline and experimental models are increased.

Figure 6: Wildfire images contained within the combined dataset. These images provide a diverse set of wildfire features for the training of the baseline and experimental models

Figure 7: Non-wildfire images contained within the combined dataset. By providing a diverse set of non-wildfire forests in different terrains, the baseline and experimental models learn to not mistake forest fire features, like rising smoke and orange fire, with common forest environments, i.e. adjacent rivers with haze and rocky mountains looming over orange hills

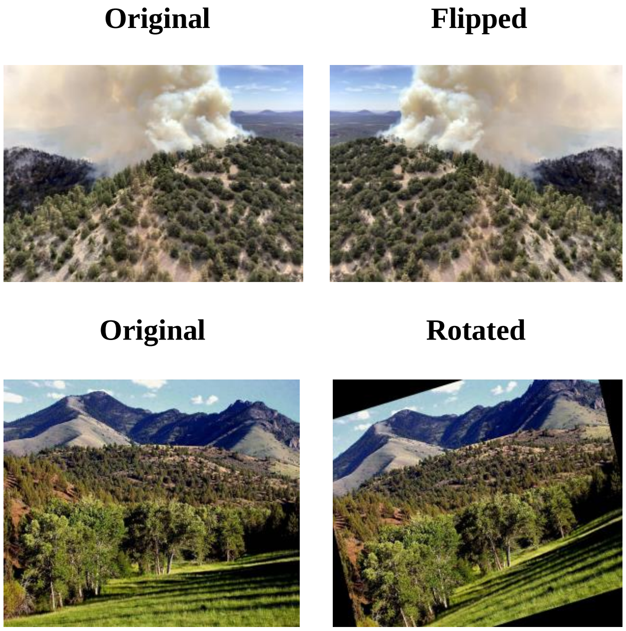
4.1. Image Compression

Due to the dataset used in this report being the combination of four different datasets, the images contained in the dataset do not have the same resolution sizes. Additionally, with some images having very large resolution sizes, the training time of both the baseline CNN model and the experimental transformer model are quite substantial. As a result, the dataset's images are resized to be 224 x 224 in resolution size. This ensures consistently sized images that are not too large are used for training both models.

4.2. Dataset Augmentation

Deep learning models, including both CNNs and transformers, require large amounts of labeled data to perform well. However, by providing a substantial amount of training data to these models, they are susceptible to overfitting to the dataset.

Data augmentation helps increase the generalizability of deep learning models to prevent overfitting. This approach involves applying different transformations to a dataset's images, maintaining their original labels, and adding them to the dataset alongside the original images. The dataset used in this report contains horizontal flipped images and 15-degree rotations that are both derived from the original dataset before such transformations.



Original **Flipped**

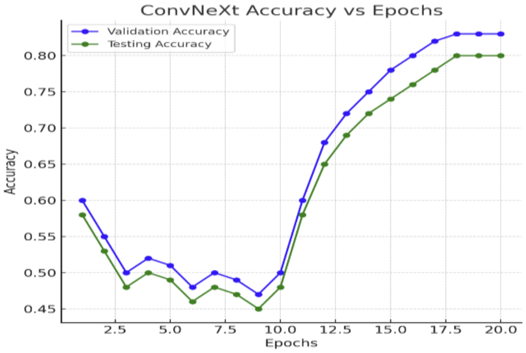
Original **Rotated**

Figure 8: Augmented images from the dataset. The top row contains a fire image before and after being horizontally flipped. The bottom row contains a no fire image before and after being rotated 15-degrees.

5. Experiments

5.1. Experiment Results

The following are the results of the baseline ConvNeXt model.



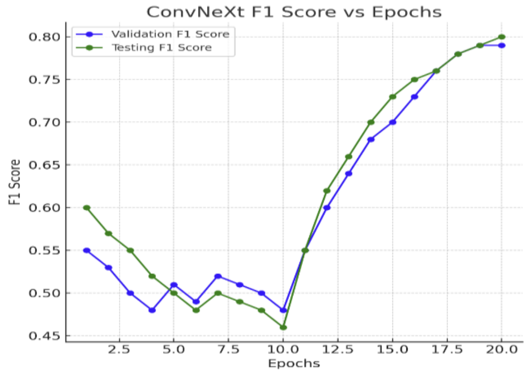


Figure 9: ConvNeXt Model Accuracy vs Epochs & F1 Score vs Epochs plots

For both plots, the y-axis represents the performance metric, which is either accuracy or F1 score, and the x-axis represents the number of epochs. As can be seen in the plots, both performance metrics jump considerably higher after around 10 epochs. This indicates that the transfer learning that takes place after 10 epochs for the ConvNeXt model significantly increases its performance.

In addition to the results of the ConvNeXt model, the following are the results of the Swin Transformer model.

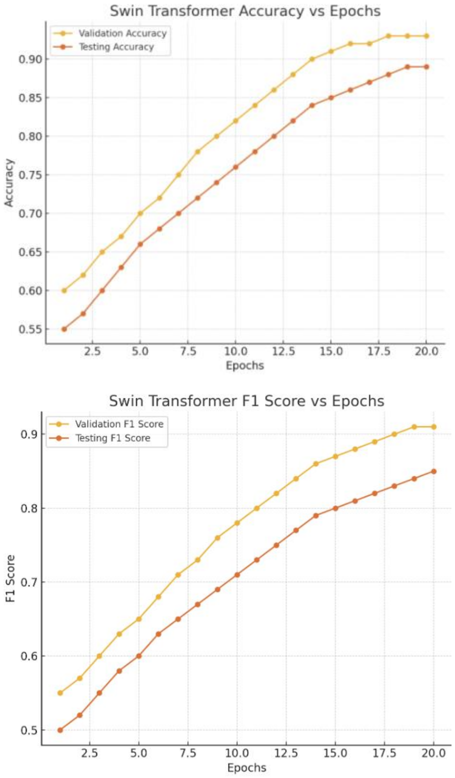


Figure 10: Swin Transformer Accuracy vs Epochs & F1 Score vs Epochs plots

Unlike the ConvNeXt model, the Swin Transformer model displays a consistent improvement in both accuracy and F1 score at every epoch. There is no considerable jump in both performance metrics after 10 epochs.

These results are likely due to the design of both models. As stated earlier, the baseline ConvNeXt model is only trained on its last layer after 10 epochs. Consequently, the pre-trained ConvNeXt model does not learn from the fire and no fire images present in the dataset within the first 10 epochs. This explains the model's poor downward trend performance for the first 10 epochs.

Once the ConvNeXt model begins transfer learning after epoch 10, the ConvNeXt model performs significantly better than previous epochs with an upward trend indicating that increasing the number of epoch iterations from 20 will lead to the ConvNeXt model to continuously improve.

In regard to the experimental Swin Transformer model, the results show the Swin Transformer consistently improving as the number of epochs increases. This performance is explained by the Swin Transformer continuously training its last layer through transfer learning for every epoch unlike the baseline ConvNeXt model.

The results shown for both models follow the expectation. That being having one model transfer learn immediately will likely allow that model to perform much stronger than another model that is only allowed to transfer learn much later than the first model.

However, there is a very interesting observation seen in the performance scores of the ConvNeXt model: the noticeable downward trend in performance of the ConvNeXt model early in the training process before transfer learning begins. The reason for this behavior is due to the ensemble learning architecture being implemented exclusively for the ConvNeXt model.

Ensemble learning was implemented only for the baseline ConvNeXt model to make up for its transfer learning deficiency up until epoch 10. The intention was for ensemble learning to improve on the baseline ConvNeXt model's generalizability and accuracy.

Instead, by combining the results of three different ConvNeXt models as one ensembled ConvNeXt model, any misclassifications or inaccuracies caused from not allowing any of the three ConvNext models transfer learn until epoch 10 are exacerbated. As a result, the baseline ConvNeXt model's performance is not flat early on as expected, but instead there is a significant poor downward performance before epoch 10 is reached.

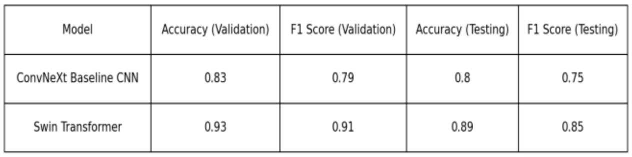
Despite the cons attributed to integrating ensemble learning to the baseline ConvNeXt model, there is positive merit in implementing ensemble learning to the ConvNeXt model evident in its performance plots.

After epoch 10, both the ConvNeXt model and Swin Transformer model see positive upward performances. However, the gaps for both performance metrics between the validation set and testing set of the Swin Transformer model begin to enlarge as the number of epochs increase. This is not the case for the ConvNeXt model as the validation set and testing set performance metrics stay very close to one another as the number of epochs grows. As a result, there is evidence of overfitting for the Swin Transformer model and not for the ConvNeXt model.

This behavior is most likely attributed to ensemble learning being integrated into only the baseline model. By averaging the results of three different ConvNeXt models, the ensembled ConvNeXt model corrects any misclassifications of one of the three ConvNeXt models. This is only possible with the ensembled ConvNeXt model being able to train itself with the intended wildfire dataset, which is why we see this positive behavior only after epoch 10.

With the ConvNeXt model's upward positive performance trend and lack of overfitting unlike the Swin Transformer, there is strong support for the ConvNeXt model being the stronger wildfire detection classifier than the Swin Transformer model, which is seemingly not the case at first glance of the performances of the two models. Thus, the intended goal of improving the generalizability of the baseline model through ensemble learning is eventually realized for the ConvNeXt model despite its early pitfalls during its training process.

The following are the final performance scores of the two models at epoch 20.



Model	Accuracy (Validation)	F1 Score (Validation)	Accuracy (Testing)	F1 Score (Testing)
ConvNeXt Baseline CNN	0.83	0.79	0.8	0.75
Swin Transformer	0.93	0.91	0.89	0.85

Figure 11: Accuracy and F1 scores for both the baseline ConvNeXt model and the experimental Swin Transformer model at epoch 20

5.2. Attention Heatmap Qualitative Analysis

Although both models achieve relatively strong performances, the two models do misclassify some non-fire images as containing fire.

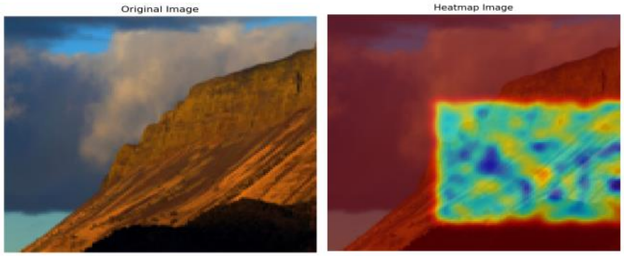


Figure 12: Example of the ConvNeXt model misclassifying a non-fire orange mountain image as a fire image

As seen above, the baseline ConvNeXt model classifies an orange mountain image as containing fire despite no fire being present in the image. This behavior is due to the original image containing abundant orange color that the ConvNeXt model interprets as a feature of wildfires. The model's behavior is indicated by the attention heatmap of the same image seen on the right.

Another misclassification is seen below.

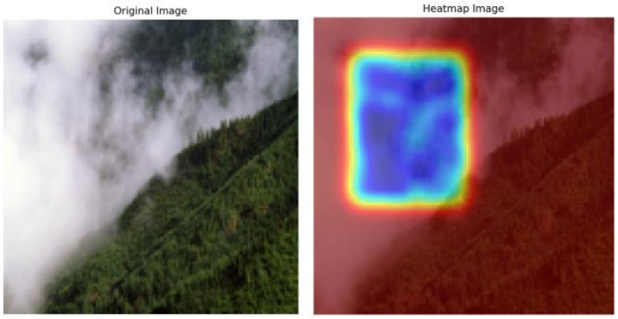


Figure 13: Example of the Swin Transformer model misclassifying a non-fire foggy forest image as a fire image

This misclassification is due to the immense amount of fog present in the original image. The Swin Transformer misinterprets the fog as smoke emanating from a wildfire as seen in the attention heatmap leading to the misclassification.

These misclassifications are due to both models misinterpreting common fire features to be present in images without fires. Adding significant dropout to both models targeted toward very common fire features, like red and orange colors as well as collection of grey mass, during the training process will likely remedy this issue.

6. Conclusion and Future Work

This report delved into the applying convolutional neural networks and transformers toward detecting wildfires. After analyzing many previous works targeted toward wildfire detection, the techniques of a few papers were integrated in building the baseline CNN and experimental transformer used in this paper. Methods to prevent overfitting were implemented like ensemble learning and data augmentation which proved to be useful in the results of the two models. Qualitative analysis was conducted using attention heatmaps to analyze possible improvements to both models.

Expanding on this work in the future would be integrating the models into a system of drones. The drones would survey dry regions likely to experience wildfires during hot seasons. Applying these models to the real-world will help provide a low cost solution to wildfires.

7. Contributions and Acknowledgements

For this project, I contributed to all parts of the report and code used to generate the results seen in this report. I would like to thank my project mentor Cem Gökmen for his guidance throughout the formulation, development, and refinement of this project, and I would like to thank Nikil Ravi for providing feedback during the implementation of my project.

.

8. References

[1] Ghali R, Akhloufi MA. Deep Learning Approaches for Wildland Fires Remote Sensing: Classification, Detection, and Segmentation. Remote Sensing. 2023; 15(7):1821. https://doi.org/10.3390/rs15071821.

[2] Saima Majid, Fayadh Alenezi, Sarfaraz Masood, Musheer Ahmad, Emine Selda Gündüz, Kemal Polat, Attention based CNN model for fire detection and localization in real-world images, Expert Systems with Applications, Volume 189, 2022, 116114, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.116114.

[3] C. Kyrkou and T. Theocharides, EmergencyNet: Efficient Aerial Image Classification for Drone-Based Emergency Monitoring Using Atrous Convolutional Feature Fusion, in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 1687-1699, 2020, doi: 10.1109/JSTARS.2020.2969809.

[4] Hong, Z., Hamdan, E., Zhao, Y., Ye, T., Pan, H., & Cetin, A. E. (2024). Wildfire Detection Via Transfer Learning: A Survey. Signal, Image and Video Processing, 18(1), 207-214.

[5] Hong-zhou Ai, Dong Han, Xin-zhi Wang, Quan-yi Liu, Yue Wang, Meng-yue Li, Pei Zhu, Early fire detection technology based on improved transformers in aircraft cargo compartments, Journal of Safety Science and Resilience, Volume 5, Issue 2, 2024, Pages 194-203, ISSN 2666-4496, https://doi.org/10.1016/j.jnlssr.2024.03.003.

[6] Coen, J.L., Schroeder, W., & Rudlosky, S.D. (2018). Transforming Wildfire Detection and Prediction Using New and Underused Sensor and Data Sources Integrated with Modeling. Handbook of Dynamic Data Driven Applications Systems.

[7] Ghali R, Akhloufi MA, Mseddi WS. Deep Learning and Transformer Approaches for UAV-Based Wildfire Detection and Segmentation. Sensors (Basel). 2022 Mar 3; 22(5):1977. doi: 10.3390/s22051977.

[8] Ghali R, Akhloufi MA, Jmal M, Souidene Mseddi W, Attia R. Wildfire Segmentation Using Deep Vision Transformers. Remote Sensing. 2021; 13(17):3527. https://doi.org/10.3390/rs13173527

[9] Z. Liu, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 9992-10002. doi: 10.1109/ICCV48922.2021.00986

[10] Rekavandi, Aref & Rashidi, Shima & Boussaid, Farid & Hoefs, Stephen & Akbas, Emre & Bennamoun, Mohammed. (2023). Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art.

[11] Saied, A. (2020). FIRE Dataset. Kaggle. https://www.kaggle.com/datasets/phylake1337/fire-dataset

[12] Dincer, B. (2020). Wildfire Detection Image Data. Kaggle. https://www.kaggle.com/datasets/brsdincer/wildfire-detection-image-data

[13] Jafar, A. I. B. (2020). FlameVision. Kaggle. https://www.kaggle.com/datasets/anamibnjafar0/flamevision

[14] Madafri, I. E. (2020). The wildfire dataset. Kaggle. https://www.kaggle.com/datasets/elmadafri/the-wildfire-dataset