Illinois Institute of Technology

Final Report:

Wildfire Detection from RBG Images using CNN, Transformer, and Hybrid Models

Matiss M Mednis, Nathan Loh, Nikolaus Schultze

CS577 - Deep Learning

Prof. Yan Yan

December 1, 2023

Introduction and Problem Description

Wildfires are detrimental to ecosystems and humanity as a whole. They can be started for various reasons and spread very quickly. Recently, with the Canadian wildfires, we have seen the global impact a large-scale forest fire can cause. Wildfires can spread rapidly and become extremely difficult to control. Forest agencies attempt to do their best to prevent them, but even the slightest initial spark can have massive consequences. Prevention is one half of the equation, and the other is detection. Early detection systems of wildfires can help prevent large-scale disasters and allow authorities to identify and respond quickly to wildfires before they become out of control. Various automated methods of monitoring wildfires exist, such as inferred cameras, lasers, and classifying RGB video feeds or images. RGB images and video from monitoring stations above the treeline may be most interesting to monitoring wildfires as RGB cameras are the most accessible and the cheapest. Satellite image detection may only be effective when fires are already at a large scale. Therefore, RGB images on the ground from monitoring stations throughout the most prone portions of a forest may also be of interest for their potential early detection capabilities. Augmented with direct human monitoring, an automated system could monitor complex places for humans and thus expand the monitoring potential beyond what humans alone could practically achieve. Once potential fires are flagged, they can be further investigated by forest services and prevented earlier. Therefore, developing models that can most accurately identify a wildfire's beginning and intermediate stages based on live RGB image feeds would be of great interest and importance in keeping ecosystems and forests intact and saving lives and property.

Challenges with wildfire detection based on purely colored images are that naturally occurring phenomena such as mist, clouds, sunsets, and sunrises can all prove difficult for image classification systems, as under proper conditions, the colors and shapes produced by these occurrences can look similar to wildfires. In those cases, approaches such as satellite monitoring or infrared detection would likely better reduce false positives. Satellite monitoring stations offer larger fields of view and an angle from above, which may help cut through minor weather patterns such as mist or a very orange and red sunset or morning sunrise. Thermal monitoring offers information on the temperature and thus would be less likely to produce false positives where the sky or mist may give off an orange or red hue similar to a wildfire.

I. Summary of previous work

Most of the work on wildfire recognition systems is very new and published in 2023. Current works typically utilize CNN MobileNetV2 models to detect wildfires [1] [2]. The data used comes from satellite images [2] [3], and in another case, uses an on-the-ground, RGB set of images [1] with MobileNetV2. In other work, we see the use of transformer models on satellite

images which produces better results than the CNN models [4]. In [1], they create a multi-level classification that first determines if there is smoke or not in an image and then further classifies whether there is a fire. This assisted in classifying images more accurately that have very red sunsets or mists in them. The separation of more granular classes was a good method that proved to do better than the binary class model. We propose training transformer models and hybrid CNN and transformer models on the on-the-ground RGB images of forest fires in various stages to improve the efficacy of camera monitoring stations in forests. We also experiment by utilizing images of fires outside of the forest fire setting to improve the model's ability to recognize the traits of fire and smoke and potentially build a multi-task learning approach pipeline with additional training images of fires and smoke in general settings. By removing the context of the forests, we hope the models might pick up on the more general traits of a fire and thus improve their efficacy in identifying the traits of fires in a forest.

II. Data Set Description:

The wildfire dataset is a comprehensive collection that explores the effectiveness of RGB imagery for forest fire detection using machine learning techniques. It comprises 2,700 aerial and ground-based images from various online sources, including government databases, Flickr, and Unsplash. These images cover various environmental conditions, forest types, geographical locations, and the intricate dynamics of forest ecosystems and fire events, making them valuable for forest fire detection research. All images are sourced from the Public Domain. The data source link provides detailed information about each image's origin URLs and resolutions. A unique feature of this dataset is utilizing a Multi-Task Learning framework designed to enhance forest fire detection by addressing multi-class confounding elements. This approach aims to improve model accuracy and reduce false alarms, especially compared to traditional classification techniques. We aim to potentially implement this framework as time permits. The dataset is divided into three main directories: Training (70%), Validation (15%), and Testing (15%). The "nofire" folder contains 1653 images, and the "fire" folder contains 1047 images. Thus about 61% of the data is labeled nofire, and 39% of the images are labeled fire. One of the challenges of wildfire data is there aren't as many available images of the fire class as obtaining images of wildfires is challenging and doesn't often occur. Within the nofire folder are 847 images containing Forested areas without confounding elements, 336 images of Fire confounding elements, and 471 images of Smoke confounding elements. Within the fire folder are 662 images of Smoke from fires and 384 images of smoke and fire. This further segmentation of the data and confounding elements may be implemented to improve model performance. The original dataset can be found at the following source https://www.kaggle.com/datasets/elmadafri/the-wildfire-dataset.

We additionally used the 'fire' labeled images of the fire dataset found at https://www.kaggle.com/datasets/phylake1337/fire-dataset to conduct our experiment on the

impact of appending additional fire images to the training dataset on model performance. 755 images labeled ‘fire’ in this dataset were added to our training dataset for the experiment. The images in the folder show fires in buildings and smoke rising from non-wildfire backgrounds.

III. Summary of Results

Algorithm	Simple CNN	Complex CNN	Pre-trained CNN	Vision Transformer (ViT)	Hybrid CNN+Transformer
Accuracy	84.63%	86.34%	88.54%	97.06%	95.61%
Precision	84.00%	85.00%	90.00%	97.09%	96.29%
Recall	92.00%	94.00%	92.00%	97.06%	94.57%
F1-Score	88.00%	89.00%	91.00%	97.07%	95.30%

Table 1. Final results of each architecture’s best model on the Wildfire test set.

In summary, our exploration into three distinct architectural paradigms—CNNs, Transformers, and a hybrid amalgamation—provided valuable insights into the landscape of wildfire detection. The progression from simple to complex CNN designs demonstrated an incremental boost in accuracy, indicating the influence of model complexity. Vision Transformer (ViT) emerged as a robust contender, showcasing consistent performance and adaptability to varying datasets. While holding promise, the hybrid model revealed the nuances associated with integrating new data sources, signaling the need for careful adjustments and extended training durations. Overall, our results underscore the efficacy of transformers and hybrid architectures in image classification tasks while emphasizing the ongoing refinement and exploration necessary to realize their full potential in real-world wildfire detection scenarios.

Methods

In this paper, we utilize and compare CNN architectures, hybrid CNN and transformer architectures, and Hugging Face’s ViT model to approach wildfire detection in RGB images on the wildfire dataset. We collect and compare all models’ accuracy, F1, recall, and precision metrics. For the ViT and hybrid models, we additionally train them on additional fire dataset images and analyze their final performance on the test set with that additional data in training. Below, we provide a detailed overview of each model implemented.

I. Simple CNN model description

The Simple CNN model comprises three convolutional layers, each followed by max-pooling, effectively capturing hierarchical features from the input images. The model contains ReLU activations in the fully connected layers, including a sigmoid activation in the output layer, and is fine-tuned with the Adam optimizer using binary cross entropy as the loss function.

II. Complex CNN model description

The Complex CNN model contains four convolutional layers with progressive max-pooling, incorporating ReLU activations in its fully connected layers. It introduces dropout layers (50% and 30% dropout rates) for regularization, with a sigmoid activation in the output layer. The model is optimized using the Adam optimizer with binary cross entropy as the loss function.

III. Pre-trained CNN model description

The pre-trained CNN model seamlessly utilizes a pre-trained VGG16 with 'imagenet' weights. Allowing the pre-trained layers to remain fixed, a flattened layer connects to a dense layer comprising 128 units with a ReLU activation. The final layer, which tailors the model for binary classification, contains a sigmoid activation—optimized with the Adam optimizer.

IV. ViT model description

For our ViT model, we utilize Google's Hugging Face ViT-base-path16-224 model, whose full description can be found at https://huggingface.co/google/vit-base-patch16-224. It is pre-trained on ImageNet-21k at a resolution of 224x224. It was first introduced in the paper *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* We leverage the power of the pre-trained model and fine-tune the model on our datasets. The images are rescaled and transformed according to the image preprocessing steps completed on the pre-trained dataset. Images are rescaled to 224x224 and normalized with a mean and standard deviation of .5 across the RGB channels. We found in the dataset that 12 images across training, validation, and test had 4 channels and were encoded as RGBA images; therefore, those were removed to successfully preprocess the data for the model.

Our best F1 score on the validation set with only wildfire images was produced with a learning rate of 2e-4, batch-size of 16, and image resizing of 384x384. In some fine-tuning tasks, resizing the images to 384x384 improves performance which was the case for this task as it increased our performance on the validation set to an F1 score when compared to resizing the

images to 224x224. We then used the same model architecture and parameters to train on the altered dataset, including the images from the fire dataset. The optimal models were found by selecting the best performance on the validation set. The results section below shows the final detailed results of the validation and test set. Each model was trained for 4 epochs with evaluation on the validation set at every 50 steps in order to find the best model that gave the highest F1 score on the validation set.

V. Hybrid model description

The hybrid model is composed of a CNN and a Transformer. However, we differ from this with the ViT model by incorporating a more complex CNN layer. We developed a similar hybrid structure to the Convolutional vision Transformer (CvT) architecture, which can be accessed from the site, https://huggingface.co/docs/transformers/model_doc/cvt. The components of this structure are a pre-trained CNN and a transformer section based around a ViT architecture. After resizing it to 224 by 224, we fed the image to the pre-trained CNN of EfficientNetB0, which we got from the Keras application library. For the transformer section, we incorporated the patches with size 32 by 32 to capture local features of the images in a 32 by 32 pixel size. We apply this to multi-head attention to allow the model to capture the complex relationships between distant image regions. Since the dataset consists of two classes, we use the binary cross-entropy loss function. In addition, we implement a learning rate scheduler to update the learning rate as it trains. The model will reduce the learning rate based on the validation loss value.

We train this on two separate datasets; however, the model's architecture is the same for both. Therefore, this section aims to compare how training additional images may impact the results without further hypertuning the model. We also evaluated the models on different testing sets to see if the new images could be accurately labeled.

Results

Below, we show the appropriate tables and graphs for the metrics of the models on the test set, as well as the validation and training losses. We also include confusion matrices for the CNN and hybrid models.

I. Simple CNN

Results Table:

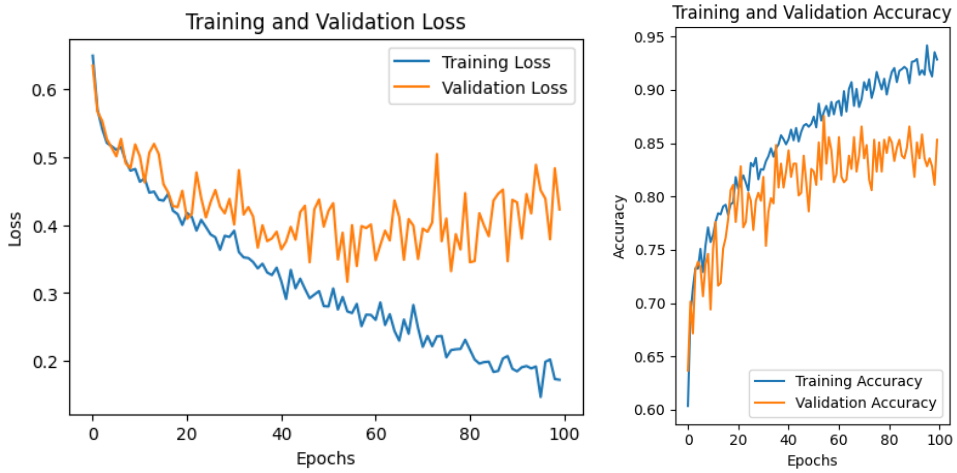
Number of Epochs	20	100
Accuracy	77.56%	84.63%
Precision	84.00%	84.00%

Recall	78.00%	92.00%
F1-Score	81.00%	88.00%

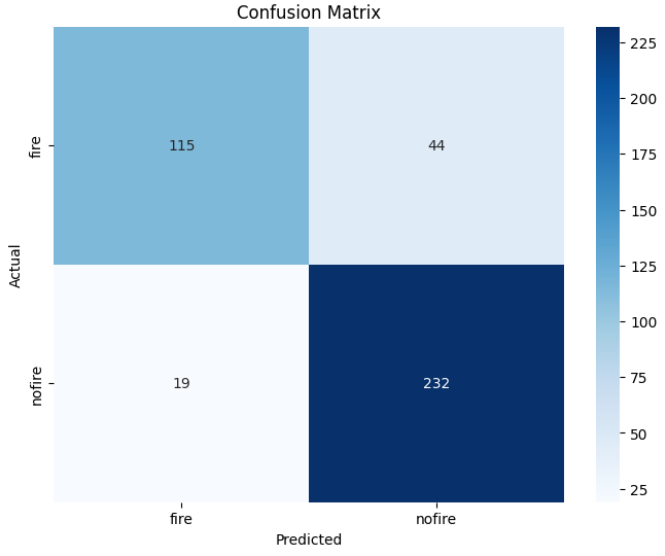
Table 2. Metrics of simple CNN on test set for given number of epochs

Training and Validation Loss and Accuracy and Confusion Matrix:

Simple CNN (100 Epochs)



Figures 1 and 2. The graphs represent the training and validation loss and accuracy for the simple CNN design trained on the Wildfire dataset. The model was trained using 100 epochs.



TP: 115
FN: 44
FP: 19
TN: 232

Figure 3. The confusion matrix illustrates the simple CNN model's results based on the Wildfire dataset.

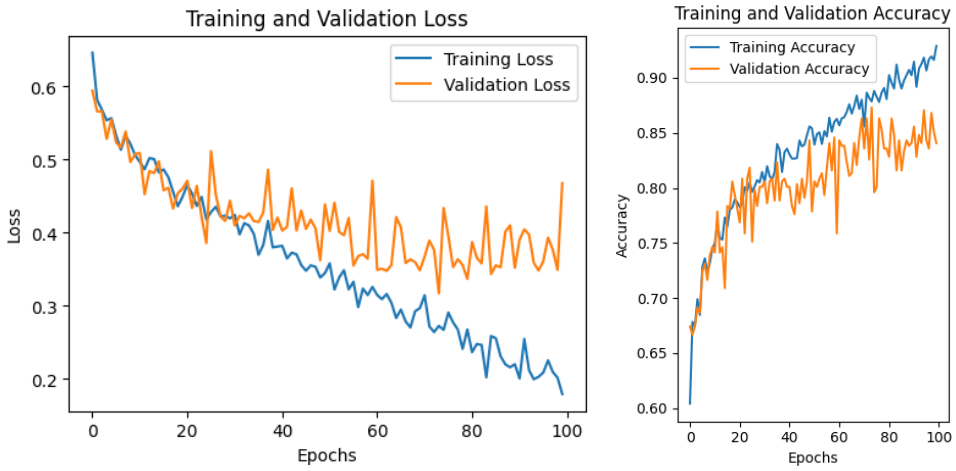
II. Complex CNN

Results Table:

Number of Epochs	20	100
Accuracy	79.51%	86.34%
Precision	79.00%	85.00%
Recall	91.00%	94.00%
F1-Score	84.00%	89.00%

Table 3. Metrics of Complex CNN on test set for given number of epochs

Training and Validation Loss and Accuracy and Confusion Matrix:

Complex CNN (100 Epochs)	 *Figures 4 and 5. The graphs represent the training and validation loss and accuracy for the complex CNN design trained on the Wildfire dataset. The model was trained using 100 epochs.*

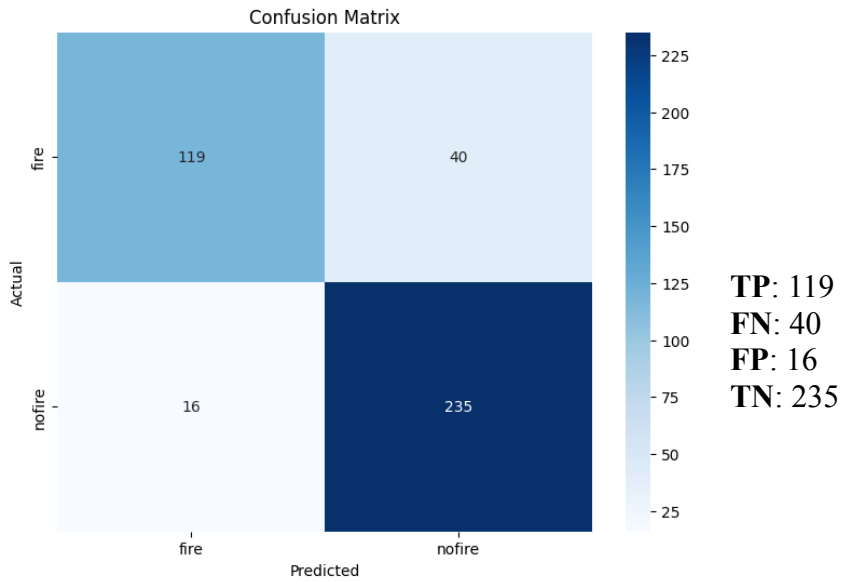


Figure 6. The confusion matrix illustrates the complex CNN model's results based on the Wildfire dataset.

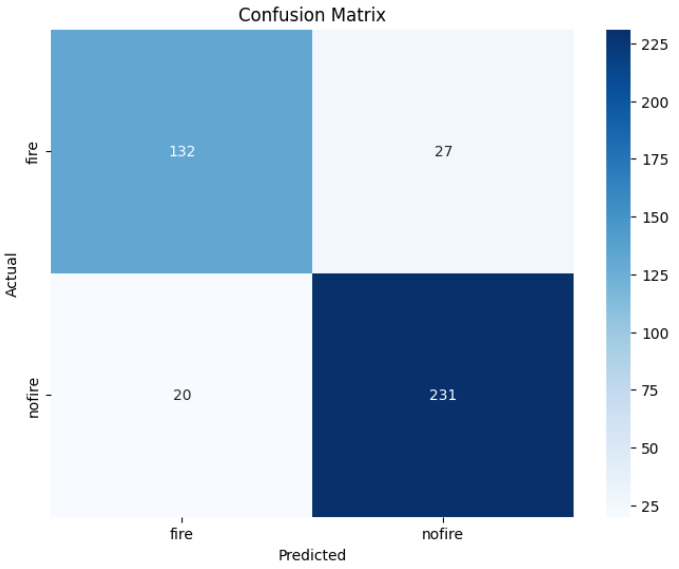
III. Pre-trained CNN

Results Table:

Number of Epochs	3	20
Accuracy	86.34%	88.54%
Precision	93.00%	90.00%
Recall	84.00%	92.00%
F1-Score	88.00%	91.00%

Table 4. Metrics of Pre-trained CNN on test set for given number of epochs

Training and Validation Loss and Accuracy and Confusion Matrix:

Pre-trained CNN (100 Epochs)	 *Figures 7 and 8. The graphs represent the training and validation loss and accuracy for the pre-trained CNN design trained on the Wildfire dataset. The model was trained using 20 epochs.*
	 **TP**: 132 **FN**: 27 **FP**: 20 **TN**: 231 *Figure 9. The confusion matrix illustrates the pre-trained CNN model's results based on the Wildfire dataset.*

IV. Vision Transformer (ViT)

Training Set	Original Data (wildfire only)	Mixed Data (fire and wildfire)
Testing Set	**Original Data**	**Original Data**
Accuracy	97.06%	96.32%
Precision	97.09%	96.35%
Recall	97.06%	96.32%
F1-Score	97.07%	96.33%

Table 5. Results of the ViT model tuned to the validation set on the Wildfire dataset test set. The Training set row represents which data set was used for training. The original data set represents the model, which is only trained on the Wildfire dataset, as the Mixed data set contains images from both the Wildfire and Fire datasets found in the Dataset Description portion of this paper. Both trained models have the same parameters and preprocessing and are tested and validated on the Wildfire dataset.

Step	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
50	0.291600	0.412709	0.824121	0.853393	0.824121	0.810793
100	0.367600	0.312459	0.856784	0.866947	0.856784	0.851129
150	0.091800	0.258709	0.919598	0.925008	0.919598	0.917856
200	0.133300	0.191740	0.927136	0.929468	0.927136	0.927575
250	0.072600	0.238450	0.929648	0.929517	0.929648	0.929558
300	0.095700	0.229176	0.934673	0.936418	0.934673	0.933888
350	0.041500	0.255763	0.932161	0.934868	0.932161	0.931161
400	0.007700	0.232388	0.932161	0.932022	0.932161	0.931935
450	0.040000	0.241171	0.934673	0.935024	0.934673	0.934213
500	0.007000	0.217023	0.934673	0.935024	0.934673	0.934213
550	0.004700	0.183596	0.952261	0.952333	0.952261	0.952291
600	0.017000	0.194442	0.942211	0.942283	0.942211	0.941935
650	0.003600	0.195863	0.944724	0.944735	0.944724	0.944500

Figure 10. Training and validation metrics on the mixed dataset. The metrics shown are on the validation set.

[401/472 49:57 < 08:53, 0.13 it/s, Epoch 3.39/4]

Step	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
50	0.311200	0.239414	0.896985	0.897887	0.896985	0.897290
100	0.255600	0.399261	0.836683	0.867690	0.836683	0.824606
150	0.126700	0.263461	0.914573	0.918694	0.914573	0.912898
200	0.138600	0.203880	0.924623	0.924817	0.924623	0.924092
250	0.073300	0.198863	0.927136	0.928317	0.927136	0.927430
300	0.040300	0.224413	0.934673	0.934598	0.934673	0.934409
350	0.034000	0.224931	0.944724	0.944724	0.944724	0.944724

Figure 11. Training and validation loss on original Wildfire dataset. (Note the model crashed in the last training run, so the epoch was not completed, but the model was saved, and thus the highest validation F1 was achieved right before it crashed.) The metrics shown are on the validation set.

V. Hybrid CNN+Transformer

Results Table:

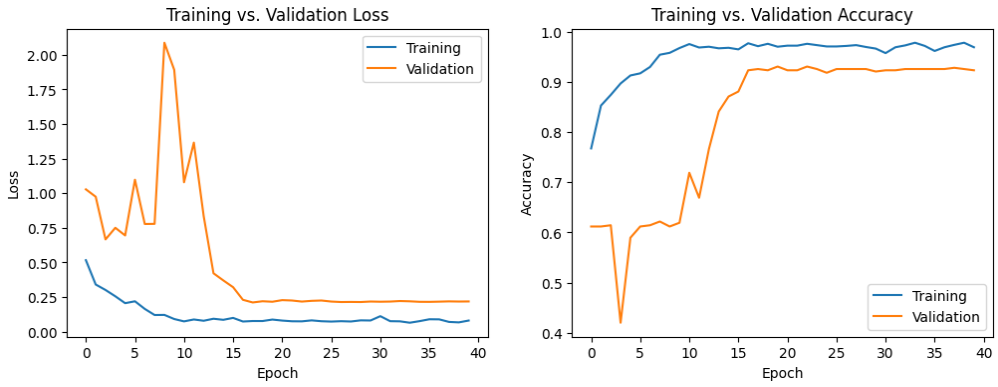
Training Set	Original Data			Mixed Data
Testing Set	**Original Data**	**Mixed Data**	**New Data**	**Original Data**
Number of Epochs	40	40	40	20
Accuracy	95.61%	83.68%	78.78%	95.37%
Precision	96.29%	83.40%	75.57%	96.11%
Recall	94.57%	86.63%	83.88%	94.26%
F1-Score	95.30%	83.26%	76.20%	95.03%

Table 6. There are two models developed using the same architecture. The first model is trained using only the original images, and the second model is based on a combination of the original and new images. For the testing set, there are three variations. The original data set comprises only the first Kaggle dataset, the mixed data set includes both data sets, and the new data set only contains the second Kaggle dataset.

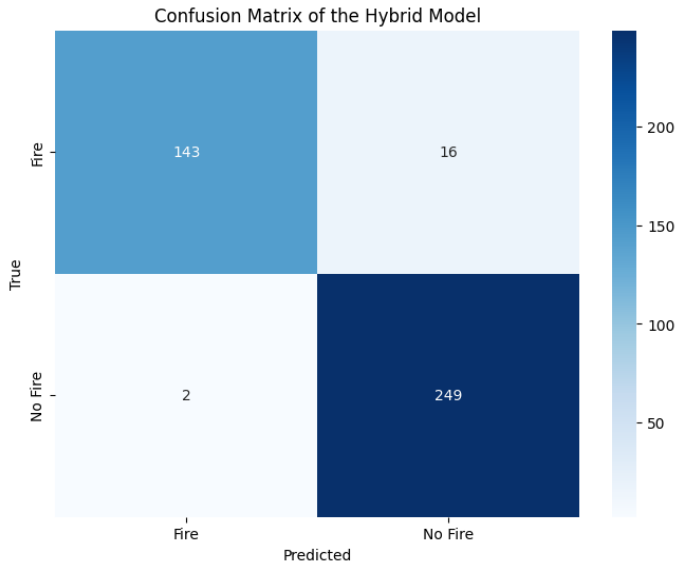
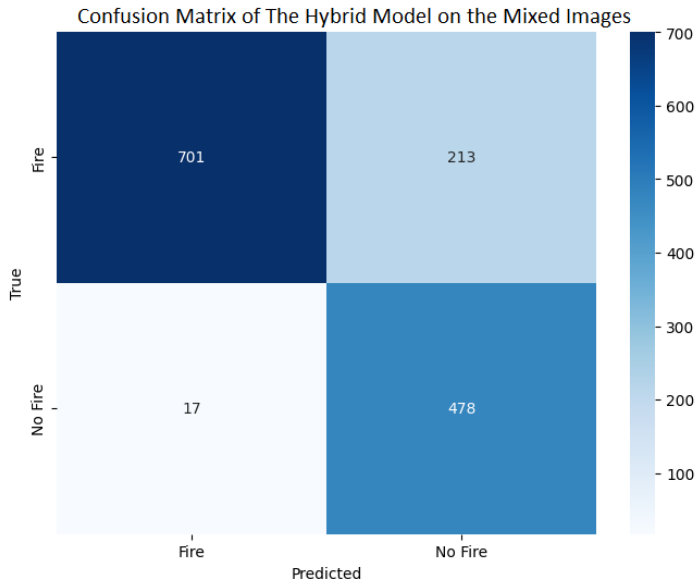
Training and Validation Loss and Accuracy and Confusion Matrix:

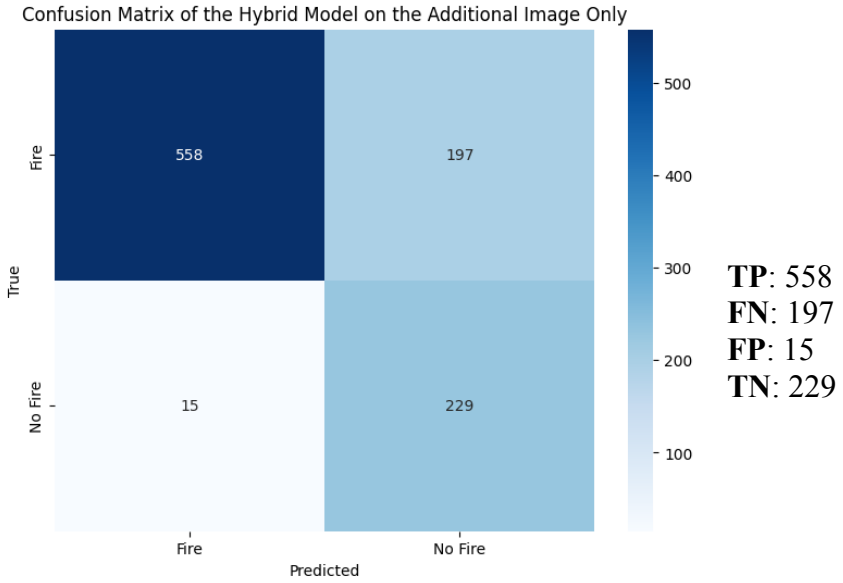
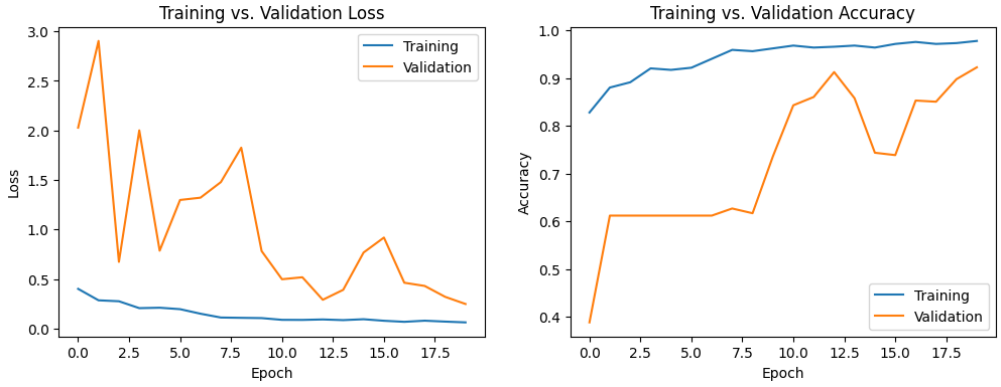
Training: Original

Testing: Original



Figures 12 and 13. The graphs represent the training and validation loss and accuracy for the hybrid design trained on the Wildfire dataset. This model was trained using 40 epochs.

	 **TP**: 143 **FN**: 16 **FP**: 2 **TN**: 249 *Figure 14. The confusion matrix illustrates the hybrid model's results based on the Wildfire dataset.*
Training: Original **Testing: Mixed**	 **TP**: 701 **FN**: 213 **FP**: 17 **TN**: 478 *Figure 15. The confusion matrix illustrates the results of the hybrid model based on a mix of the Wildfire dataset and the FIRE dataset.*

Training: Original **Testing: New**	 **TP**: 558 **FN**: 197 **FP**: 15 **TN**: 229 *Figure 16. The confusion matrix illustrates the hybrid model's results based on only the FIRE dataset.*
Training: Mixed **Testing: Original**	 *Figures 17 and 18. The graphs represent the training and validation loss and accuracy for the hybrid design trained on the Wildfire dataset and FIRE dataset. This model was trained using 20 epochs.*

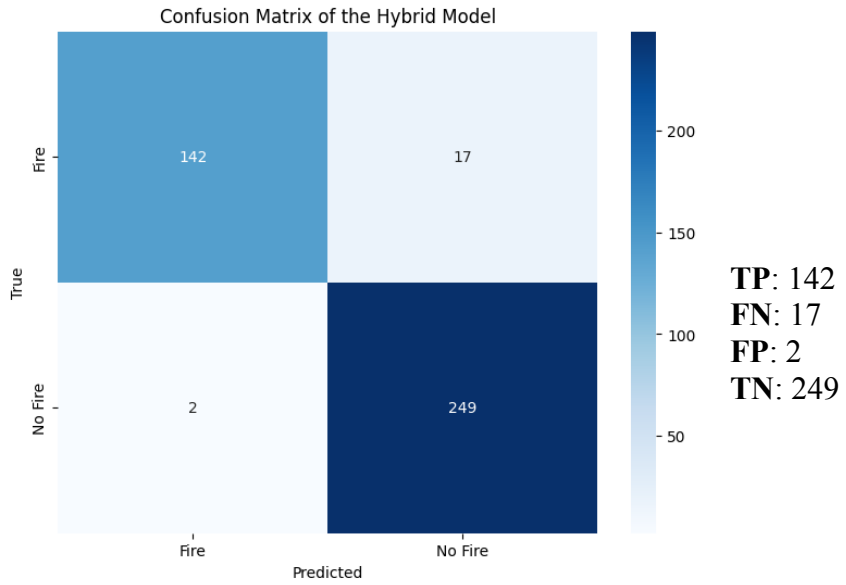


Figure 19. The confusion matrix illustrates the mixed hybrid model's results based on the Wildfire dataset.

Final Comparison of best models the Wildfire Test Set

Algorithm	Simple CNN	Complex CNN	Pre-trained CNN	Vision Transformer (ViT)	Hybrid CNN+Transformer
Accuracy	84.63%	86.34%	88.54%	97.06%	95.61%
Precision	84.00%	85.00%	90.00%	97.09%	96.29%
Recall	92.00%	94.00%	92.00%	97.06%	94.57%
F1-Score	88.00%	89.00%	91.00%	97.07%	95.30%

Table 7. The best results of all models are put together. We compared each model's results and listed their accuracy, precision, recall, and F1 scores. As noted, there are significant improvements from CNNs to Transformers.

Discussion of Results:

Regarding the project, we wanted to compare three main architectures: CNN, Transformers, and a hybrid between the two. Due to the complexity of this task and the limited time, we chose to work on three levels of CNN designs, a pre-trained transformer, and a hybrid design incorporating pre-designed models. For all the models, we evaluated them under four main categories: accuracy, precision, recall, and F1 score. However, it is essential to note that the best models of each architecture are based on the validation loss and accuracy rather than the testing set results. In addition, the best models do not indicate the architectures' best models in general for wildfire detection, as further hyperparameter tuning and adjustments can be applied to improve the design.

The three levels of CNN are broken down into simple, complex, and pre-trained designs. As noted in *Table 7*, the accuracies improve by around 2 percent as we continue to increase the complexity of the model, starting at 84.63% to 86.34%, and finally 88.54%. Each of the CNN models has higher recall percentages compared to the precision. This indicates that CNNs tend to incorrectly predict no fires when there are actually fires (FN) compared to predicting no fires when they are truly labeled fires (FP). However, this can also be caused by the disparities between fire and no-fire images. In the data set, there are more no-fire images than fire images. When analyzing each CNN model's training and validation loss and accuracies, they fluctuate up and down. This could potentially be fixed by utilizing a learning rate scheduler to reduce the learning rate as it trains, allowing the model to converge smoothly. We will refer to the pre-trained CNN results to compare the CNN design with the other main architectures.

For the ViT model, we see generally consistent performance across recall, precision, and accuracy metrics. It notably does not tend to have a clear imbalance between FP and FN rates. This is consistent across both the original wildfire dataset and the mixed dataset, including fire and wildfire images. This is shown in *Table 5,* as the F1 score is within +- .03% of all other metrics for both models. We see the validation set performance increase with each epoch, and a greater number of epochs would possibly give even better results. Surprisingly, the test set performance is greater than the validation set performance in the original and mixed dataset-trained models. The mixed dataset model only does slightly worse on the test set than the original dataset model, but notably, the mixed dataset model reaches a higher maximum validation F1-score of 95.22% in *Figure 10* opposed to 94.47% in *Figure 11*. Therefore, one might assume the mixed data model might actually outperform the original dataset model before deployment. There is certainly room to do longer training for both models, as we still see the potential for the validation performance to improve with more epochs. We see an impressive ability for the model to learn from the fire images and apply that to the wildfire images to produce very high results on just wildfire images. Given that the ViT model was only

hyperparameter tuned on the original dataset, and then the same model was applied to the mixed dataset, there could be potential room to tune a model on the mixed dataset and get even better results.

We created two models and trained them on different data sets for the hybrid design of CNN and Transformer. The first model is trained off the Wildfire Dataset, while the second is trained off of a combination of that set and the FIRE Dataset. Due to the lack of computational resources (limited GPU), the second model is only trained with 20 epochs. This is significant as in *Figures 17 and 18,* the validation loss is still decreasing and the validation accuracy is increasing. This indicates that increasing the number of epochs can improve the model's results. At 20 epochs, we see that the mixed dataset does slightly worse than the first model, which was trained with 40 epochs. We also see that in *Figures 12 and 13,* the validation loss and accuracies of the first model stagnate at around 18 epochs. When evaluating the first model with the addition of the FIRE dataset, we see that it does significantly worse, having around 78.78% accuracy. This indicates that the model trained on the Wildfire dataset does not fully translate to additional images. Overall, we see that training on the new images from the FIRE dataset can help improve the model results, and that there needs to be further training and improvements.

Overall, we see a clear improvement in results when comparing the CNN and Transformer models. ViT and the hybrid model did significantly better for all metrics than CNN. This indicates that utilizing transformers or a mix between the CNN and transformers can improve image classification results. When analyzing a pure transformer and a modified transformer (one that uses a more complex CNN layer), using the pre-trained transformer for wildfire detection seems better. However, this could change based on future works to improve and adjust the models. As noted in the hybrid model discussion, there are improvements that can be made, indicating that it become just as good or better than the ViT model. Based on the ViT model, many values, such as patch size and number of attention heads, are pre-determined. However, in our hybrid model, we manually assigned these values, suggesting that more variables need to be hyper-tuned, which can lead to significant improvements in results.

Conclusions and Future Work

Conclusions

In conclusion, exploring diverse architectural frameworks for wildfire detection sheds light on critical nuances in model performance. The comparative analysis spanning simple and complex CNN designs, the Vision Transformer (ViT), and a hybrid model demonstrate the dynamic interplay between model complexity and accuracy. While CNNs demonstrated incremental improvements, ViT and the hybrid model presented better results.

The outcomes underscore the transformative potential of transformer-based architectures, as exemplified by ViT, and the synergies achieved through hybrid models. However, challenges in integrating new datasets reveal the importance of detailed adjustments and prolonged training. The study's significance lies in guiding future endeavors, encouraging further fine-tuning, exploring diverse datasets, and an extended understanding of the intricate dynamics between architectural choices and model performance. As we advance, this research provides a foundation for developing more refined and adaptable wildfire detection models, crucial for addressing real-world fire prevention and management challenges.

Future Work

Future work could involve more extensive training sessions to refine model performance for all models. Greater exposure to the dataset during training may reveal patterns in the data and contribute to the overall robustness of the models. Additionally, a thorough exploration of hyperparameter tuning and model architectures could uncover optimal configurations, presenting an opportunity for fine-tuning to enhance each model's effectiveness in fire detection tasks.

Acquiring additional computation resources is crucial to bolster the experiment's scope further. Increased computational capabilities would support the exploration of more complex neural network architectures and expedite the training process, allowing for the implementation of intricate models that may unveil latent features within the data. Moreover, future efforts should explore diverse datasets beyond the current scope, introducing new challenges and environments to assess the models' adaptability and generalization capabilities.

We can also utilize the 755 'fire '-labeled images from the 'fire' dataset available on Kaggle and retrain the Simple CNN, Complex CNN, and pre-trained CNN. These additional images, showcasing fires in buildings and smoke rising from non-wildfire backgrounds, are introduced to assess the impact of appending diverse fire scenarios to the training dataset. This augmented dataset is anticipated to yield insights into their adaptability to various fire-related contexts, contributing to a more comprehensive understanding of their performance in real-world scenarios.

Github:
https://github.com/nschultze/CS577Project