Image Augmentations For Satellite Imagery Fire Risk Assessment

Mohamed Owda
Stanford University
mohamed8@stanford.edu

Adrian Saldana
Stanford University
asaldana@stanford.edu

Abstract

Wildfires are a natural event where large swaths of land burn uncontrollably, risking the lives of flora and fauna as well as destroying property and infrastructure. There is growing interest in using deep learning methods to assess the fire risk of an area using the increasingly available corpus of remote sensing images. In this paper, we build upon the work of [16] and use their labeled Fire Risk dataset which has 70,331 to classify the fire risk of an area on a scale of 1 to 7. We initially use Vision Image Transformers (ViTs) and ResNet models pretrained on ImageNet to establish baselines. We then apply physical data augmentations, and find that ResNets do not respond to data augmentations while ViTs do respond to data augmentations and improve performance. As ViTs performed the best after augmentations, we explored further techniques to improve accuracy. We considered various learning rates, L2 regularization alone and in conjunction with augmentations, and pretraining on the training set. We achieved our highest model validation accuracy of 64.52% which outperformed [16] ViT accuracy from combining augmentations and L2 regularization. However, we recognize this model did still demonstrate overfitting, and thus recommend more work to fight overfitting, namely the procurement of a greater distribution of training data, to improve accuracy. The code for the project can be found at [1].

1. Introduction

Wildfires are a natural disaster where large portions of land burn in an uncontrolled manner. In recent times we see an upward trend of global fire activity being exacerbated by climate change [14]. This increase in fire activity poses a serious risk to society, as wildfires can risk the lives of humans and animals in the vicinity of the burning, destroy property and infrastructure. Additionally, the burning of vegetation releases stored carbon and produces smoke which can result in respiratory issues for those inhaling smoke and an increase in greenhouse gases in the atmosphere.

The improvements in remote sensing technology and increased access to these images has led to the use of satellite imagery in land use and land type deep learning classification tasks, with Convolutional Neural Networks (CNNs) being especially popular for use in these problems [13]. As a result, there is interest in using satellite imagery on deep learning tasks to predict which areas are most susceptible to wildfires. This will allow for better consideration of land management and infrastructure planning as it relates to wildfires.

[16] present the "FireRisk" dataset consisting of 91,872 labelled images across 7 fire risk classes. They also present a supervised learning benchmark performance of 63.20% classification accuracy on a ResNet model pretrained on ImageNet1k, and 63.31% on a Vision Image Transformer (ViT) pretrained on ImageNet1k. In this paper, we attempt to use the [16] dataset to improve upon the benchmark classification accuracy found by [16]. The input is a 224 x 224 image from the Fire Risk dataset. We then use ResNet and ViT models to output a fire risk class, on a scale of 1 to 7. In order to improve performance, we first explore the application of various physical data augmentations techniques in order to increase the training data and fight against overfitting. From this, we focus on improving ViT performance by optimizing the learning rate, experimenting with L2 regularization alone and alongside data augmentations, and further pretraining on downsampled 32 x 32 resolution images on the training set before running finetuning on 224 x 224 resolution images.

2. Related Work

In approaching the problem of predicting fire risk using satellite imagery, [4] used NASA's MODIS satellite images to create a time series of vegetation change, along time series data of humidity and temperature to derive a fire potential index of a region. Similarly, [15] attempted to assess wildfire danger in areas of China based on topography, weather, and potential fuel for fires, with this fuel data being based on modeling applied on top of MODIS images, and [10] also used land use features derived from satellite

[1] https://github.com/internationalmo/cs231n_final_project

imagery to train machine learning models to determine the susceptibility of a region to fire. However, we see that these approaches use non-image features to generate fire risk predictions, and the ability to generate these features requires domain knowledge. To our knowledge, [16] is the first paper to present a fire risk classification model based solely on image data.

In review of the satellite imagery deep learning literature, [3] compared existing methods and found that finetuned CNN-based models have the highest accuracy in the field with respect to image classification. [9] demonstrated the success of ViT models in satellite imagery detection tasks as well.

Data augmentation is a technique where synthetic modifications are applied to the original dataset. A few of these augmentations include random flips, random crops, and random zooms. Data augmentations are done to increase the amount of training data and make the model more robust to overfitting. In a survey of data augmentation in deep learning, [17] found that image classification accuracy improved with flipping, scaling ratio, rotation, noise injection, cropping, translation and sharpening. [1] found that specifically for satellite imagery, random flipping exhibited the greatest increase in classification accuracy, with a 6% improvement over baseline when training a VGG19 model. Shearing, zooming, and rotation also showed improvements over baseline, and these improvements over baselines held when augmentations were combined. [2] explored the value data augmentations to examples created from generative adversarial networks and baseline satellite imagery from EuroSAT. They demonstrated that random horizontal flip, random vertical flip, and random rotation caused modest improvements in accuracy (approximately 0.3%).

Pretraining is the practice of initializing the weights of a deep learning model to perform a task on a large dataset, before further finetuning the weights to perform on a downstream task. [12] first showed how CNNs pretrained on the ImageNet dataset were successful in transfer learning on downstream remote sensing tasks. Building upon this work, [6], [7] showed ViT improve performance when pretrained on in-distribution remote-sensing imagery before being finetuned on a more niche satellite imagery dataset.

3. Dataset

Our dataset is comprised of 7 fire risk classes, {1: High, 2: Low, 3: Moderate, 4: Non-burnable, 5: Very High, 6: Very Low, 7: Water}, which denote the potential of wildfire burning in the given satellite image. We have a total of 70,331 high-resolution labelled remote sensing images originally presented by [16], collected using the National Agriculture Imagery Program (NAIP). The minimum image resolution is 1 meter, and to ensure image quality, cloud cover does not comprise any more than 10% per quarter of the image patches and the sun must be at least 30 degrees above horizon at the time the image is taken. There is no snow or flood coverage in the dataset. The original images are 270 x 270 pixels and consists of three channels; R, G, and B and the pixel values lie in the range [0, 255]. Our paper retrieves data from [2] which upsizes the images to being 320 x 320. The dataset is imbalanced, with 1729 examples of the "Water" label compared to 21757 examples of the "Very Low" label.

Table 1. Example Counts

Label Name	Number of Examples
Non-burnable	21757
High	17959
Low	10705
Moderate	8617
Very Low	6296
Very High	3268
Water	1729

We split the dataset into 80% training, 10% validation, and 10% testing and stratified by the labels to keep the label proportions balanced across training, validation, and testing.

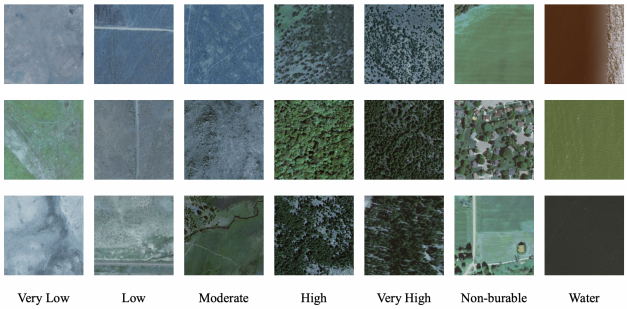

Figure 1. Examples of images in the dataset and there original labels. Graphic from [16].

4. Methods

We have a labeled dataset containing 70,331 RGB satellite images of Earth, each with a resolution of 320x320 pixels. Our goal is to maximize the accuracy of correctly identifying the fire risk classification for each input image.

4.1. ResNet Architecture

The first baseline model architecture we considered to solve the problem is a ResNet-50 which was first introduced by [8]. The original Fire Risk paper ([16]) used ResNets in their findings making the ResNet a natural choice in our paper to establish our own baselines as well to compare to the original paper.

[2] https://huggingface.co/datasets/blanchon/FireRisk

ResNet-50 is a CNN model which has 50 layers. After the initial layer of 64 7x7 convolutional filters and then a 3x3 max pool, we do 48 layers of convolutions followed by batch normalization and a ReLU activation function. The final layer is an average pool followed by a fully connected mapping. Skip connections are also incorporated between layers in order to allow for effective training of such a deep network. As there are 7 fire risk classes, we use the multi-class cross entropy loss function for the ResNet-50

$$\mathcal{L} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} y_{ij}\log(p_{ij})$$

where N is the batch size, C is the number of classes, $y_{ij} = 1$ when i is the true class for example j and $y_{ij} = 0$ otherwise, and p_{ij} is the probability we assign to class i of example j.

The ResNet-50 model we use was pretrained on ImageNet-1k allowing for us to do transfer learning on the pretrained weights. We made the decision to not train the model from scratch since we have a relatively modest 56,284 examples available for training. With this amount of examples, especially as the data is imbalanced, we were concerned that training a deep network would result in overfitting and poor performance. Furthermore the original FireRisk paper also used a pretrained ResNet and thus using a pretrained model in our paper allowed for a more similar comparison. We do recognize that satellite imagery is quite different than ImageNet's dataset, throwing into question the benefit pretraining brings and address this in Pretraining. We used the ResNet model from [3].

4.2. Vision Image Transformers Architecture

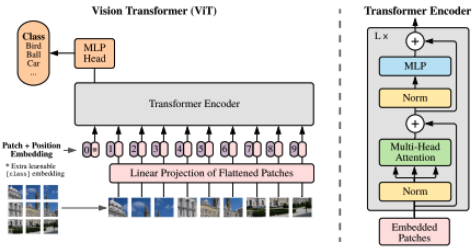

Figure 2. ViT overview from [5]

The other baseline model architecture we considered to solve the problem is a Vision Image Transformer (ViT) which was introduced by [5]. The original Fire Risk paper ([16]) used ViTs in their results making the ViT another logical selection to establish our own baselines as well to compare to the original paper and to the ResNet performance.

In contrast to the ResNet, the ViT divides the image into 16 x 16 patches. Every patch is then compressed into a vector by concatenating the pixel values across the RGB channels then projecting this vector down to a $\mathbb{R}^{786}$-dimensional input vector. A positional encoding is also added to the vector to communicate the position of the patch in the image. There are 12 hidden layers in the encoder. In each layer we alternate between multi-head attention and MLP block. There are 12 attention heads in the multi-head attention, with layernorm applied before every attention and MLP sublayer and residual connections after every block. Again a Cross-Entropy loss is used.

We used the ViT from [4] which was pretrained on ImageNet-21k at 224 x 224 resolution.

4.3. Data Augmentations

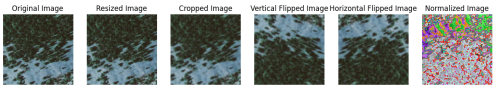

Figure 3. Image Processing Pipeline

In many real-world applications, labeled satellite imagery to assess the fire risk of a region may not be readily-accessible, forcing researchers to make use of small datasets. This small training data may result in model-overfitting and the model not being robust enough to handle small differences in imagery caused by lighting or cloud cover. Thus, a large area we were interested in exploring was data augmentations on our data, which allows for a more diverse set of training examples to fight against over fitting. In this paper, we focused on 4 main data augmentation strategies:

- **Random Horizontal Flipping**: We apply a random horizontal flip of the image with a probability of 0.5.
- **Random Vertical Flipping**: We apply a random vertical flip of the image with a probability of 0.5.
- **Normalization**: We normalize the images around a mean and standard deviation. The mean and standard deviation come from the ImageNet dataset that the respective ResNet and ViT models were pretrained on.
- **Random Crop**: Apply a crop to a random portion of the image and scale it back to 224 x 224 (the input resolution of the image).

[3] https://huggingface.co/docs/transformers/main/en/model_doc/resnet

[4] https://huggingface.co/docs/transformers/en/model_doc/vit

Normalization, vertical flipping, horizontal flipping, and cropping are marked as N, V, H, and C for brevity. These data augmentations were applied either individually or in coordination with one another. The transformations are ordered such that V will always precede H, N, C if any of V, H, N, or C occur, and H always precedes N and C if any of H, N or C occur, and N always precedes C if either of N or C occur. This ordering was chosen arbitrarily, and future work can look towards the significance of the ordering of the transformations. The functionality for the data augmentations comes from PyTorch.

4.4. Pretraining

In addition to data augmentation, we recognize that the ImageNet dataset is largely geared for natural images and tasks more oriented for natural images such as object detection. ImageNet does not have a large representation of satellite imagery and thus is not a dataset designed for tasks that are comparable to assessing the fire risk of a satellite imagery. This made us consider whether the pretrained weights, which come from being trained on ImageNet, are the best weight initializations for our tasks. In order to test whether a different approach to pretraining could be helpful, we attempted to pretrain ViTs on the training set before running a final round finetuning.

To do so, we took the original 320 x 320 images from the training set and downsampled them to being 32 x 32. This downsampling was intended to make training run faster while still giving us pretrained weights that could transfer on higher resolution images. We then trained the ViT pretrained on ImageNet-21k for a predetermined number of epochs on the downsampled images. These images did not have data augmentations applied to them. To finetune the images, we initialized new models with the weights loaded in from the pretrained models. We then trained these models on 224 x 224 resolution images again from the training set for a predetermined number of epochs.

5. Experiments/Results/Discussion

5.1. Effect of Data Augmentation on ResNet and ViT Models

In this section, we experimented on establishing baselines on a ResNet and ViT model before applying image augmentation to explore if augmentations improved validation accuracy.

The ResNet-50 model is finetuned on the training set using a learning rate of 1e-3 and a batch size of 32 trained for 5 epochs. The ViT model is finetuned on the training set using a learning rate of 2e-4 and a batch size of 32 trained for 5 epochs. The learning rates were chosen such that they did not lead to behavior indicative of exploding/vanishing gradients while improving validation accuracy with successive epochs. The epoch size of 5 was chosen as an arbitrary baseline such that we saw improvements in validation accuracy over the course of training. In Improvements on Baseline ViT we consider how the learning rates and other hyperparamters effect the performance on Vision Image Transformers specifically. The batch sizes were chosen to maximize hardware and memory usage.

Our baseline approach is only resizing the images to being 224 x 224, with a width and height of 224 chosen to be consistent with [11]. From this baseline, we augment the data through (1) normalization based on mean and standard deviations extracted from the respective models, (2) random vertical flipping with a probability of 50%, (3) random horizontal flipping with a probability of 50%, and (4) random cropping with a probability of 50%. These augmentations were then applied individually or in combination with one another. Normalization, vertical and horizontal flipping, and cropping are marked as N, V, H, and C respectively in our figures and tables.

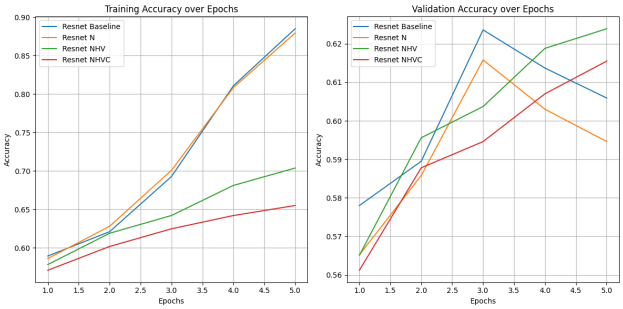

Figure 4. Resnet Training and Validation Accuracy - Image Processing

Method	Highest Validation Accuracy
Baseline	0.6236
N	.6158
NVH	.6239
NVHC	0.6155

Table 2. Resnet Validation Accuracy - Image Processing

Figure 4 shows our ResNet training and validation accuracy as we add additional techniques to our image processing pipelines. Table 2 lists our highest validation accuracy across each experiment.

As demonstrated in Figure 4, without any image augmentation, we are able to achieve a training accuracy of 88% and validation accuracy of 60.5% after 5 epochs, with the maximum validation accuracy being 62.36%. Adding normalization does not have a significant impact on training accuracy, but does slightly reduce our validation accuracy. Vertical and horizontal flipping have the intended regularizing effects, as training accuracy decreases to low 70% and validation accuracy increases by several percentage points

by epoch 5. Based on the plots, random flipping and cropping could benefit from continued training as training and validation accuracy have not yet plateaued.

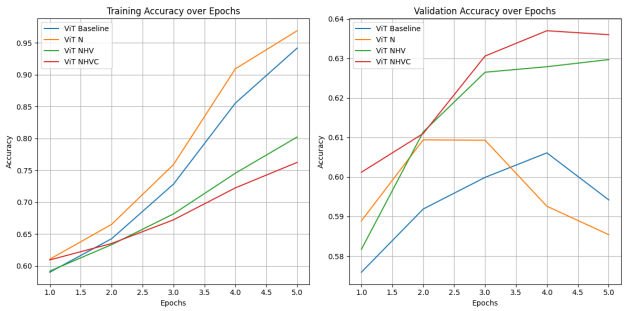

Figure 5. ViT Training and Validation Accuracy - Image Processing

Method	Highest Validation Accuracy
Baseline	0.6061
N	0.6094
NVH	0.6297
NVHC	0.6370

Table 3. ViT Validation Accuracy - Image Processing

The ViT demonstrates similar trends as the ResNet model. Figure 5 and Table 3 show our experimental results for the ViT model. Normalization added a modest improvement in training and validation accuracy. Horizontal and vertical flipping performed the intended regularization as training accuracy decreases to between 75% and 80%, and validation accuracy increases when compared to our baseline. Similar to ResNet, we could see improved training and validation accuracy from continued training with vertical and horizontal flipping augmentations.

The primary metric of classification we chose was accuracy as this was also done in [16]. Figure 4 shows a confusion matrix of our best performing ViT model. Although the validation accuracy is only 63.7%, anecdotally we see many of the errors are closely adjacent to the true classification, which represents that when the model is ”wrong” it still is helpful can be helpful in giving some sense of the fire risk.

From our results, we see that data augmentations do little to help ResNets, with all different data augmentation regimes having a similar accuracy to each other and to the baseline. The result of the baseline is around 62.36% and the NVH augmentation gives the highest validation accuracy of 62.39%, which is a negligible difference to the baseline. On the otherhand, with NVHC the ViT outperforms the baseline ViT, which had a validation accuracy of 60.61%, by 3% to have an accuracy of 63.7%.

From our graphs, we see that the baseline ViT has a training accuracy of close to 94% while the baseline ResNet

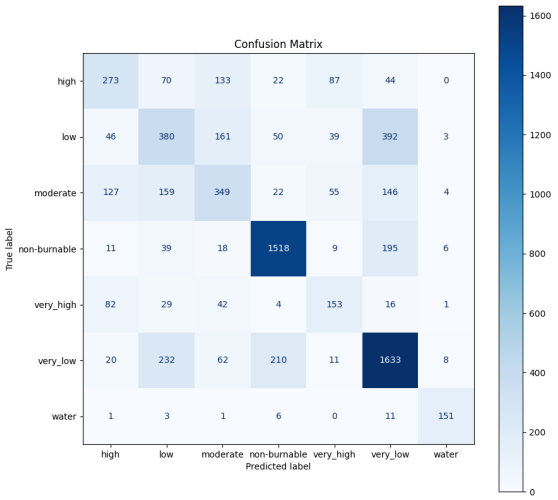

Figure 6. ViT NVHC - Confusion Matrix

has a training accuracy of 88%. This means the the ViT is more susceptible to overfitting compared to ResNet. For the ViT, the NVHC augmentation decreased training accuracy to 76% while elevating validation accuracy to 63.7%. Thus, we see that the data augmentations improve performance by fighting against overfitting. But, even with data augmentations the discrepancy between training and validation accuracy points to more work needed to prevent against overfitting, such as regularization techniques and novel training data. We explore the notion of greater regularization techniques in Improvements on Baseline ViT.

While the ResNet also demonstrates overfitting, in contrast the data augmentations for the ResNet did not help against this overfitting in a meaningful way in the final results. Thus, to imrpove ResNet accuracy, data augmentations should be seen as secondary improvements, while more emphasis should be placed on other regularization approaches and an increase in training data.

5.2. Improvements on Baseline ViT

After realizing ViT was best performer after data augmentations, we worked through various techniques to see if we could elevate the ViT performance. We explored the impact of adjusting the learning rate, L2 regularization, combining augmentation with L2 regularization, and pretraining on downsampled input images. Additionally we included F1 score to give an additional evaluation metric, though for many cases this simplify followed the trend of the validation

accuracy.

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

First, we experimented with adjusting the learning rate. We started with a general sweep across orders of magnitude, 1e-2 to 1e-5 as seen in Table 4. Learning rates on the order of 1e-4 to 1e-5 performed the best, achieving a validation accuracy of 61.62% and 61.28% respectively. A plot of our training and validation accuracy and F1 score can be seen across 5 epochs in Figure 7. We then did a fine-grained sweep within this learning rate range. Table 5 show these results. Additionally, we limited our training two epochs to save on compute time. During our fine-grained learning rate sweep, we achieved a validation accuracy 62.76% with a learning rate of 2e-4 and 1e-4.

Learning Rate	Highest Validation Accuracy	F1 Score
1e-2	0.3093	0.0675
1e-3	0.5346	0.4391
1e-4	0.6162	0.5846
1e-5	0.6128	0.5655

Table 4. Highest Validation Acc and F1 Score - Magnitude LR Sweep

Learning Rate	Highest Validation Accuracy	F1 Score
3e-4	0.6146	0.5592
2e-4	0.6276	0.5755
1e-4	0.6276	0.5790
5e-5	0.6232	0.5766
3e-5	0.6193	0.5659
1e-5	0.5996	0.5369
8e-6	0.5930	0.5145

Table 5. Highest Validation Acc and F1 Score - Granular LR Sweep

Based on these experiments, the optimal learning rate is between 1e-4 and 1e-5. A learning rate of 1e-4, as shown in Figure 7, demonstrates overfitting: the training accuracy approaches 95%, but the validation accuracy remains in the low 60s. In contrast, a learning rate of 1e-5 is too low. The training accuracy plateaus at 68% after 5 epochs, with a validation accuracy also in the low 60s. These experiments informed our decision to continue using learning rate of 2e-4 as additional experiments could help control overfitting.

Next, we explored the impact of L2 regularization by varying the weight decay parameter, as shown in Table 6 and Figure 8. L2 regularization penalizes large weights in the loss function, encouraging the model to favor smaller and more distributed weights. Smaller weights reduce sensitivity to the input and can improve generalization to unseen data. It is expected that as weight decay increases,

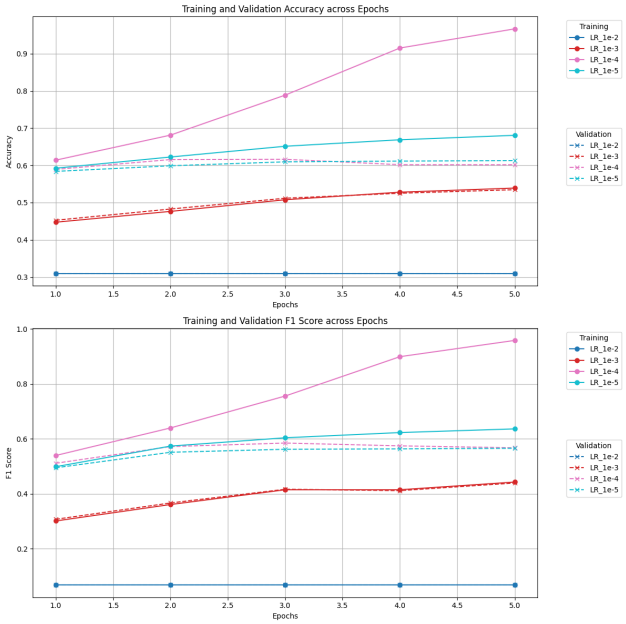

Figure 7. Validation Acc and F1 across Epochs - LR Sweep

training accuracy should decrease, thereby reducing overfitting and improving generalization. Figure 8 shows that as weight decay increases, training accuracy decreases, with weight decay values above 0.1 having the most impact. A large weight decay of 10 impedes our model's ability to learn. Ultimately, the utilization of L2 regularization and the reduction in training accuracy did not translate to significant improvements in model generalization as validation accuracy remains similar.

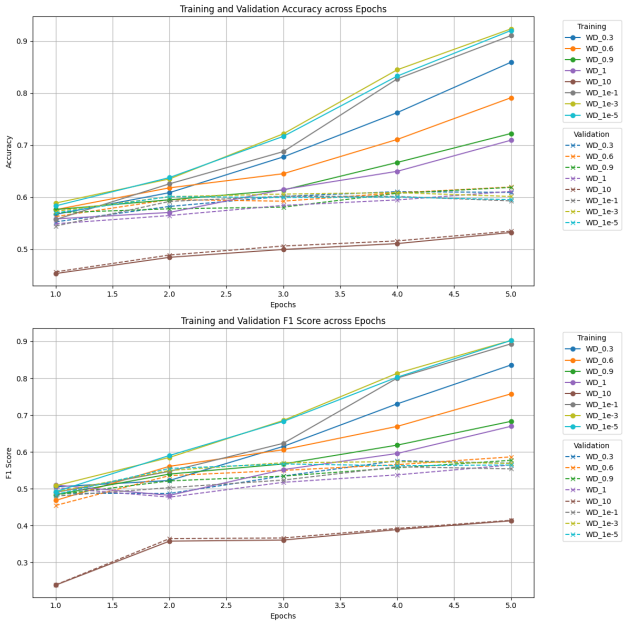

Figure 8. Weight Decay Sweep

Weight Decay	Highest Validation Accuracy	F1 Score
10	0.5347	0.4144
1	0.6111	0.5646
0.9	0.6187	0.5772
0.6	0.6195	0.5865
0.3	0.6105	0.5761
0.1	0.6105	0.5694
1e-1	0.6017	0.5597
1e-3	0.6087	0.5736
1e-5	0.6010	0.5674

Table 6. Weight Decay Sweep

For the third experiment, we wanted to see if we could improve upon our best ViT augmentation results from section 5.1 by adding L2 regularization. With all augmentations applied (NVHC), the model still seemed to be overfitting by the 5th epoch as the training accuracy continued to increase with a stable slope while validation accuracy began decreasing. In addition to augmentation, we added some weight decay as shown in Figure 9 and Table 7. A combination of image augmentation (NVHC) and L2 regularization (WD=0.001) gave us the best result of our experiments with a validation accuracy of 64.52%. Additionally, by the 5th epoch, validation accuracy for augmentations NVHC with weight decay .01 or .001 had not yet begun to decrease, demonstrating potential gains with further training. In terms of creating a highly generalizable model, adding several regularization techniques ensures no single feature is over emphasized.

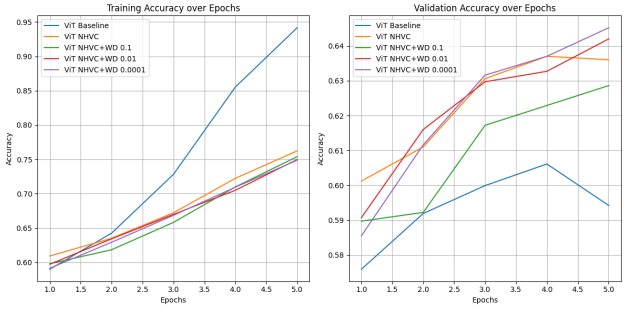

Figure 9. ViT Training and Validation Accuracy Image Processing and Weight Decay

Method	Highest Validation Accuracy
Baseline	0.6061
NVHC+WD$_{0.1}$	0.6286
NVHC+WD$_{0.01}$	0.6420
NVHC+WD$_{0.0001}$	0.6452

Table 7. ViT validation accuracy with image processing and weight decay

Lastly, we explored the impact of pretraining our model on a downsampled version of our input image and then fine-

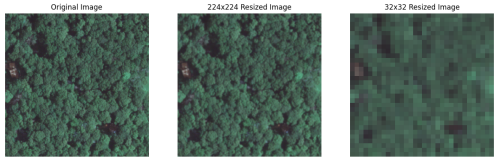

Figure 10. Resized Examples

tuning on the standard 224x224 images. During the pretraining step, we evaluated the models performance while varying the learning rate shown in column "Pretrained Val Acc" in Table 8. Similar to our previous learning rate experiment, a learning rate of 1e-4 performed the best, giving us a pretrained validation accuracy of 60.18% on 32x32 input images. Next, we finetuned the model on our 224x224 input images shown in column "Finetuned Val Acc" in Table 8 and Figure 12. Regardless of the pretraining learning rate or pretrained validation accuracy, there is no improvement in the final finetuned model. For example, using learnings rates 1e-3 and 1e-4 result in the pretrained validation accuracy of 44.76% and 60.18% respectively. The best finetuned validation accuracy that utilized those pretrained weights is 60.18% and 60.20% demonstrating that pretraining had no significangt impact.The expectation was that pretraining would allow us to improve convergence and achieve a higher overall validation accuracy, but this was not the case. It is likely that the features learned in the 32x32 image are not transferable to the full 224x224 image. Figure 1 shows how different the downsampled images are from each other which could explain why pretraining doesn't improve our accuracy. Since we did not see any benefit from pretraining, we decided not to continue experimenting with additional augmentations in conjunction with pretraining. Although pretraining did not improve our validation accuracy, an interesting finding is that a training on a 32x32 downsampled image performs similarly to training on higher resolution images.

Pretraining LR	Pretrained Val Acc	Finetuned Val Acc
1e-3	0.4476	0.6018
1e-4	0.6018	0.6020
5e-5	0.5528	0.6044
1e-5	0.5556	0.6018
1e-6	0.5074	0.59960

Table 8. Downsampled Validation Accuracy and F1 Score

6. Conclusion and Future Work

In this paper we utilized deep learning computer vision techniques to classify fire risk using satellite imagery. We first demonstrated the baseline performances of ResNet and ViT models and how they exhibited signs of overfitting.

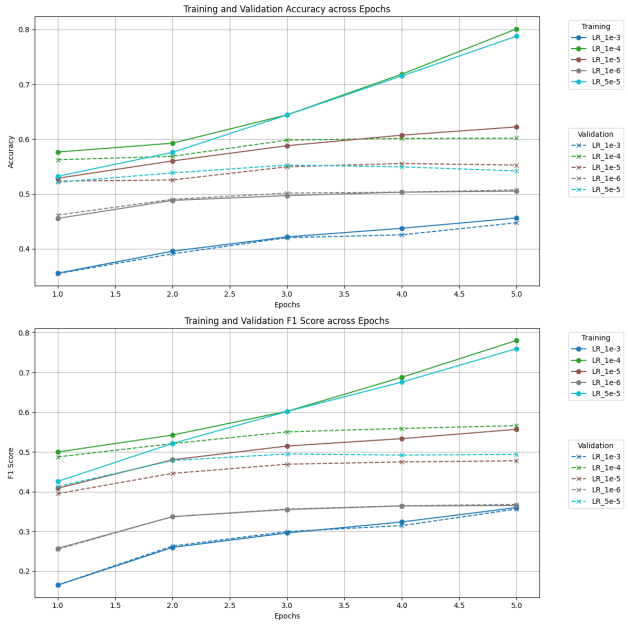

Figure 11. Downsampled Training and Validation Accuracy

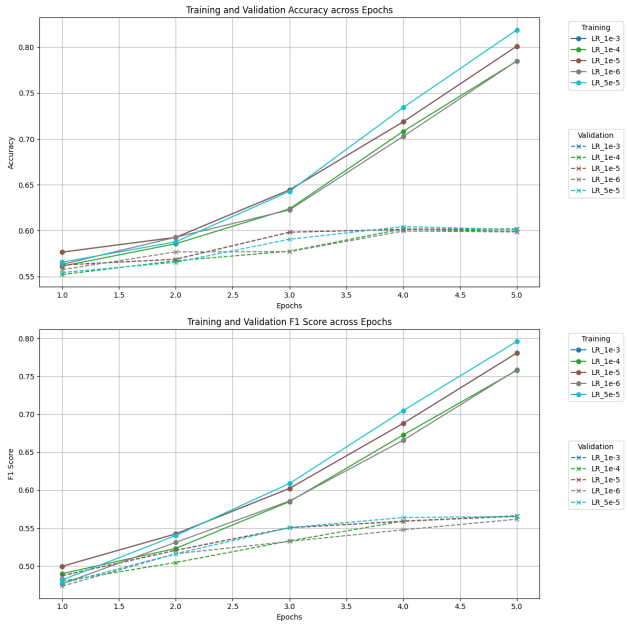

Figure 12. Pretrained on Downsampled Training and Validation Accuracy

We then explored data augmentations as a means to prevent against overfitting. While ViTs did respond to data augmentations by improvements in validation accuracy over the baseline, ResNets did not respond. We then further explored improvements to the ViT models. For our experiments, first, we determined our optimal learning rate to be used with ViT by doing a comprehensive learning rate sweep. Next, we explored L2 regularization alone, but this did not improve the generalization or validation accuracy of our model. We then applied in combination L2 regularization with data augmentations, and we were able to achieve a maximum validation accuracy of 64.52%. Lastly, we experimented with pretraining on downsampled input images and finetuning on the 224x224 images. Pretraining on downsampled images did not improve our overall validation accuracy, but it demonstrated that a downsampled image of 32x32 may be sufficient for classifying fire risk.

Future work may explore the use of highly compressed, downsampled satellite imagery for fire risk classification. Utilizing downsampled images would greatly reduce the computation required for training and inference. This would allow for more effective hyperparameter searching. The reduced computation could also enable realtime fire risk classification. Additionally, as the best performing model continued to demonstrate overfitting, more work can be put towards gathering more training data to have the model see a wider range of image features.

References

[1] M. Abdelhack. A comparison of data augmentation techniques in training deep neural networks for satellite image classification. *ArXiv*, abs/2003.13502, 2020.

[2] O. Adedeji, P. Owoade, O. Ajayi, and O. Arowolo. Image augmentation for satellite images, 2022.

[3] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[4] G. H. de Almeida Pereira, A. M. Fusioka, B. T. Nassu, and R. Minetto. Active fire detection in landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:171–186, Aug. 2021.

[5] 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, 2021.

[6] A. Fuller, K. Millard, and J. Green. Satvit: Pre-training transformers for earth observation. *IEEE Geoscience and Remote Sensing Letters*, PP:1–1, 01 2022.

[7] A. Fuller, K. Millard, and J. R. Green. Transfer learning with pretrained remote sensing transformers, 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[9] J. Horváth, S. Baireddy, H. Hao, D. M. Montserrat, and E. J. Delp. Manipulation detection in satellite images using vision transformer, 2021.

[10] B. Kalantar, N. Ueda, M. O. Idrees, S. Janizadeh, K. Ahmadi, and F. Shabani. Forest fire susceptibility prediction based on machine learning models with resampling algorithms on remote sensing data. *Remote Sensing*, 12(22), 2020.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In

F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[12] D. Marmanis, M. Datcu, T. Esch, and U. Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

[13] R. Naushad, T. Kaur, and E. Ghaderpour. Deep transfer learning for land use and land cover classification: A comparative study. *Sensors*, 21(23):8083, Dec. 2021.

[14] I. Prapas, A. Ahuja, S. Kondylatos, I. Karasante, E. Panagiotou, L. Alonso, C. Davalas, D. Michail, N. Carvalhais, and I. Papoutsis. Deep learning for global wildfire forecasting, 2023.

[15] X. Quan, Q. Xie, B. He, K. Luo, and X. Liu. Corrigendum to: Integrating remotely sensed fuel variables into wildfire danger assessment for china. *International Journal of Wildland Fire*, 30:822, 10 2021.

[16] S. Shen, S. Seneviratne, X. Wanyan, and M. Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning, 2023.

[17] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen. Image data augmentation for deep learning: A survey, 2023.