FireSight – Wildfire Detection through UAV Aerial Image Classification

Stanford University

Sharmila Reddy Nangi, Amrita Palaparthi

{srnangi, amritapv}@stanford.edu

Introduction

Problem

- **Over 8800 fires burned approximately 2.5 million acres in 2021 in CA.** Early fire detection is crucial to mitigate and control post-fire effects. **UAVs** collect immense data on natural disasters.

Goal:

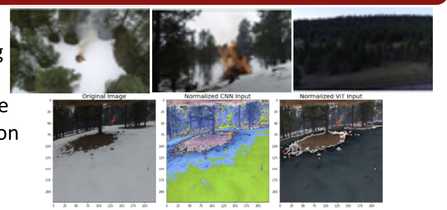
- **Detect fires from drone images with Deep Learning -** Given an input aerial image, perform binary image classification to predict one of the two classes (Fire or No-Fire).

We propose the use of **Transformers for wildfire classification**, exploring data augmentation, regularization, ensembling and compression to boost performance and support deployment onto low-memory UAVs. While prior approaches focus on CNNs for natural disaster detection and scene classification from UAV imagery, **we present an ensemble of ViT and CNN models with geometric data augmentation, achieving 82.28% accuracy on our task.**

Dataset

Dataset

- Fire Luminosity Airborne-based Machine learning Evaluation (FLAME) - aerial images collected by drones during prescribed burns in an Arizona pine forest (224x224 resolution). Test set is collected on a different type of UAV hardware to address generalizability challenges



Subset	Class	Number of Images
Train + Validation	Fire	25018
	No Fire	14357
Test	Fire	5137
	No Fire	3480

Image Normalization:

- ViT uses mean/sd of 0.5, CNN models use mean: [0.485, 0.456, 0.406], sd: [0.229, 0.224, 0.225]}

Metrics: Accuracy, F1 Score

Methods

Baselines/CNN Models

- Xception Network trained from scratch - FLAME baseline
- DenseNet and ResNet pretrained on ImageNet
  - Binary Linear Classifier trained on output features

Vision Transformer

- Pretrained on ImageNet, fine-tuned on FLAME data

Data Augmentation:

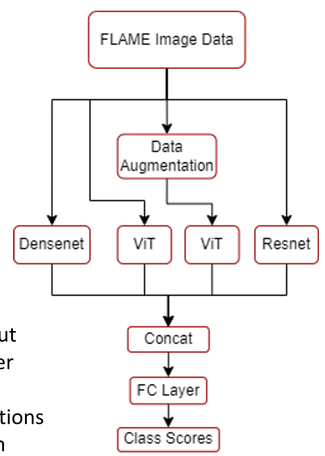
- Horizontal Flip, Rotation by ≤15°, Color jitter, Grayscale

Ensemble Strategies:

- **Voting**: Majority vote across model predictions
- **Confidence:** Model output with highest softmax probability associated with predicted class is chosen
- **Feature Concatenation:** Linear classifiers are removed, output feature maps are concatenated and passed through a FC layer

Model Compression

- Linear classifier is trained over different ViT layer representations
- Later layers of ViT are removed to examine tradeoff between model performance and model size (in MB)



Experiments and Analysis

Single Model Training:

- Fine-tuning pre-trained models leads to better performance over models trained on FLAME from scratch
- Vision Transformers show significant improvement over CNN models

Data Augmentation:

- ViT+DA accuracy indicates that geometric augmentations are effective in promoting model generalizability
- Color Jitter and Gray scale reduce model performance indicating that color is an important factor for classification
- Crop/ Cut remove or hide the presence of fire in images
- Confusion Matrix indicate lower False Negative rate.

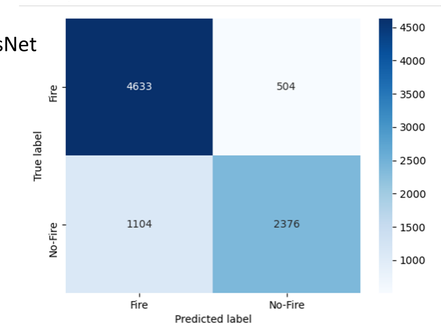
Ensemble Strategies:

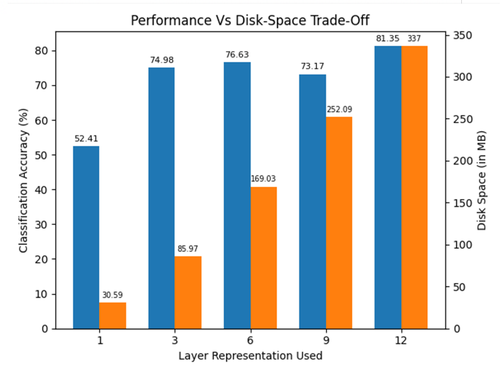
- Model trained with feature concatenation achieves the highest performance
- Best Ensemble for Concatenation Scheme - ViT, ViT+DA, ResNet, DenseNet
- Best Ensemble for Voting Scheme - ViT, ViT+DA, ResNet
- Best Ensemble for Confidence Scheme - ViT, ViT+DA, ResNet
- ViT and CNN models learn complementary features.

Model Compression:

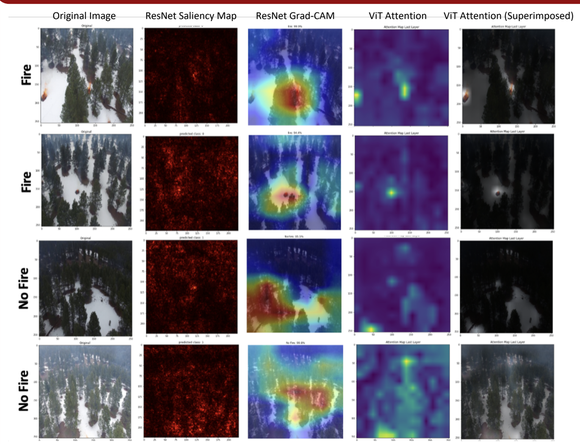
- ViT learns redundant information from layers 6-12 (visualised in the attention weights)
- Classifiers trained with features from only the first 6 ViT layers offer a 50% reduction in disk space with only a ~5% drop in model performance, indicating a diminishing payoff in performance of increasing model size at later layers

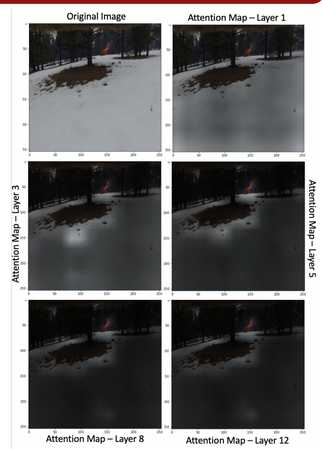
Model	Validation		Test	
	Accuracy	F1	Accuracy	F1
Xception	97.00	0.96	49.09	0.58
DenseNet	98.09	0.97	70.35	0.53
ResNet	97.69	0.97	73.20	0.61
ViT	99.95	0.99	78.25	0.68
ViT+DA	99.85	0.99	81.35	**0.75**
Ensemble-Vote	99.92	1.00	81.94	**0.75**
Ensemble-Conf	99.95	1.00	81.40	0.74
Ensemble-Concat	**100.0**	**1.00**	**82.28**	**0.75**





Qualitative Analysis





Our CNN models and ViT focus on different, but complementary areas in images - ViT concentrates on areas of fire presence, while ResNet looks at larger regions and smoke

Conclusion

Key Takeaways

- Data Augmentation enables ViT to generalize across UAV hardware, significantly outperforming the CNN baseline
- Model ensembling combines features from CNNs and ViT

Impact and Application

- Tracking wildfire perimeters, targeted UAV deployment to verify potential fires (e.g. those spotted by ALERTWildfire)

Limitations and Future Work

- Further model compression through knowledge distillation
- Augment training data to include more head-on angles and occluded fires (failure cases shown below)