Improving Deep Learning-Based Wildfire Smoke Plume Detection with a Multi-Model Ensemble Approach

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

With the increasing frequency and severity of wildfires, there is an urgent need for effective and rapid wildfire and smoke detection tools. Recent advancements in computer vision have demonstrated the potential of deep learning models, particularly neural networks, to automate the partitioning of high-resolution images into labelled segments. However, single-model approaches can struggle with generalization and accuracy in diverse conditions. To address these challenges, we propose using an ensemble of deep learning models to produce more accurate annotations of wildfire smoke plumes and their relative density (light, medium, heavy) in Geostationary Operational Environmental Satellite imagery. Our preliminary results indicate that ensemble techniques can improve performance compared to using a single model. This approach aims to provide a more reliable and accurate tool for real-time monitoring of smoke, ultimately informing fire and hazard management efforts and contributing to climate resilience and adaptation strategies.

1 Introduction

Increased wildfire activity in recent years has led to a rise in smoke and particulate matter in the atmosphere, posing greater risks of respiratory illnesses and other air quality-induced health issues [1]. Effective and timely wildfire and smoke detection tools are thus essential for supporting hazard management and mitigating risks to human health.

The National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environmental Satellites (GOES) provide high spatial and temporal resolution imagery of North America [2], which can be leveraged to detect the presence and density of smoke plumes. The NOAA Hazard Mapping System (HMS) Fire and Smoke Product currently relies on human analysts to annotate the presence of smoke over North America using GOES imagery [3]. However, this product is limited by the availability of human analysts and their time. Specifically, annotations are outputted only once to several times a day and usually have a delay between smoke occurance and the annotation. To address these limitations, we use the existing HMS dataset for training and leverage advancements in deep learning to automate the detection of smoke from GOES imagery. Deep learning models, particularly encoder-decoder neural networks, have shown promise in automating the semantic segmentation (labelling images on a pixel-wise basis with multiple classes) of high-resolution images [4]. By automating this task, we can enable more frequent detection of smoke plumes, which will inform active wildfire monitoring and impacts to air quality.

This proposal focuses on enhancing the capability of deep learning models to detect smoke through the use of multi-model ensemble techniques. It has been shown for classification tasks that ensemble

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

methods, which combine the predictions of multiple classifiers, can often perform better than a single classifier [5]. Particularly, utilizing a diverse set of classifiers in an ensemble is important to achieve the improvement in performance [6]. Furthermore, when using neural networks, combining the predictions of multiple independently-trained models can improve generalization and detection accuracy [7–9]. In this proposal, we analyze various ensemble methods for the smoke detection task.

2 Data and Methods

The dataset we use consists of 183,672 samples, each with three spectral channels (C01-C03) of GOES imagery paired with HMS smoke annotations (pixel-wise labels of smoke density of light, medium, or heavy) for a specific time and location. The data spans 2018-2024, and we use 2023 for validation and 2022 for testing, with the remaining years used for training. This ensures the testing and validation data is independent of the training data.

We utilize a variety of pre-developed encoder-decoder architectures that were designed for semantic segmentation contained within the Segmentation Models Pytorch library [10]. We select architectures that include different features such as multi-scale fields-of-view and precise boundary detection [11–13], which are important for accurately detecting smoke plumes that can vary in size. Additionally, we select the best-performing single architecture and trained it with 12 different seeds to generate different initial random weights. These models are trained independently for 24 hours on 8 Nvidia P100 GPUs using the Adam optimizer, a learning rate of 1e-3, a binary cross entropy loss function, and batch size of 128. After training, each model is selected based on its best validation Intersection over Union (IoU) score (Equation 1) which quantifies the alignment between the model prediction (y_i^*) and the ground truth (y_i).

$$\mathrm{IoU}_{\text{overall}} = \sum_{i=\text{light}}^{\text{heavy}} |y_i \cap y_i^*| \div \sum_{i=\text{light}}^{\text{heavy}} |y_i| \cup |y_i^*| \tag{1}$$

The ensemble method we are using in this preliminary analysis is an unweighted average of N model outputs [8]. A schematic of this approach is shown in Figure 1. To explore how performance improves with a variety of model combinations, we vary the number of ensemble members (1-12 models) for combinations of model architectures and initial random seeds. To our knowledge, these ensemble methods have not yet been used for wildfire smoke detection.

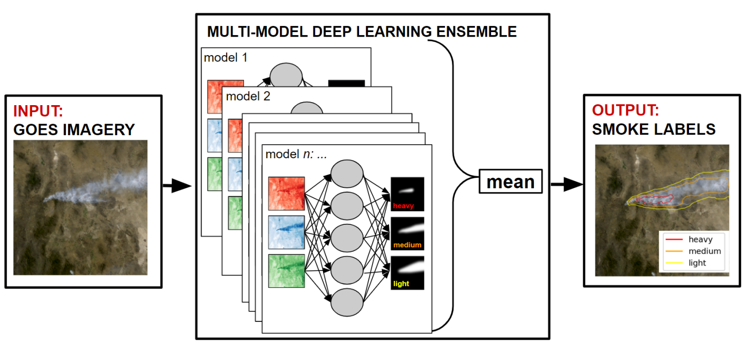

Figure 1: Multi-Model Ensemble Framework. GOES imagery is inputted to N independently-trained models whose output is combined with an unweighted average to produce the ensemble prediction of pixel-wise smoke labels.

3 Results

Table 1 shows the IoU scores for individual models and ensembles. The ensemble of 8 different architectures outperforms the individual models, with an improvement in the IoU score over all densities and for each density individually. The ensemble of 8 models (with the same architecure, PAN) with different initial weights also outperforms the individual models, with a similar improvement in the IoU scores. Figure 2 shows the IoU performance over all smoke densities as a function of ensemble size for the two ensemble schemes. The ensemble with different initial weights generally improves as models are added to the ensemble. This improvement is likely due to the different initializations leading to the models searching different parts of the parameter space and thus finding different minima of the loss function. The ensemble of different architectures improves with more models up to 8 models, but then starts to decrease in performance. This decrease in performance could be due to the additional architectures not being as well suited for the task, or the additional models not having enough variation in model bias to improve ensemble performance. Future work will aim to clarify exactly how different ensemble types and sizes reduce error and improve generalization capabilities. Figure 3 shows an example of smoke plume detection from the testing dataset. The ensemble predictions have smoother boundaries than the individual model outputs, making the prediction more comparable to the human-drawn polygon annotations.

Table 1: IoU results across three classes of smoke (light, medium, heavy) and over all densities. Presented for different individual models of different architectures (DLV3P [11]; PAN [12]), along with the archiecture ensemble and random initial weights ensemble, where N denotes the number of models in the ensemble.

	Heavy	Medium	Light	Overall
Single Model: DLV3P	0.347	0.441	0.666	0.599
Single Model: PAN	0.349	0.478	0.664	0.604
Architecture Ensemble (N=8)	0.400	0.507	0.692	0.635
Random Initial Weights Ensemble (N=8)	0.409	0.512	0.684	0.631

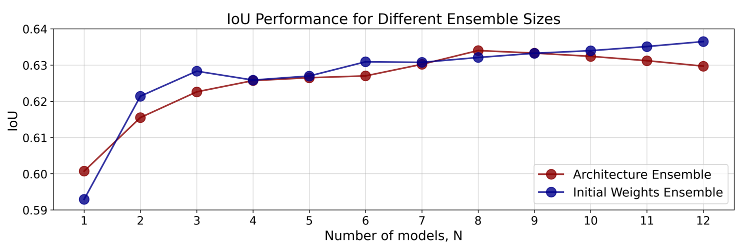

Figure 2: Ensemble IoU over all smoke densities as a function of ensemble size for two ensemble design schemes: random initial weights (blue) and architecure-based (red).

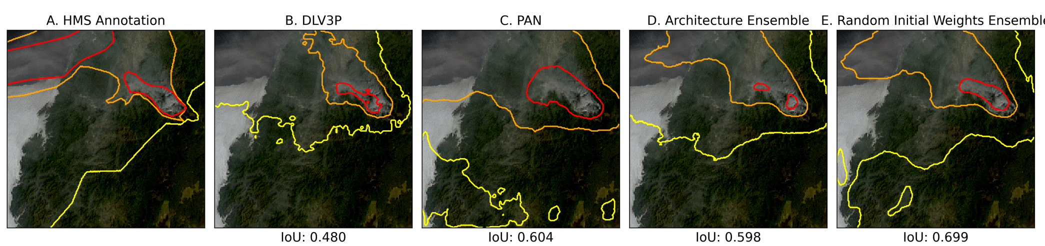

Figure 3: Example of smoke plume detection at (43.37, -123.25) on 2022/10/15 15:50 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. Panel A displays the ground truth annotation; Panels B-C show the predictions of two individual models; Panel D shows the prediction of an architecture-based ensemble (N=8); Panel E shows the prediction of an ensemble (N=8) made with models initialized with different random weights.

4 Conclusions and Future Work

This proposal explores two schemes for building ensembles of deep learning models that both improve on testing set IoU and smooth annotation boundaries. However, further investigation is required to understand why the architecture-based ensemble decreases in performance after 8 models, what the optimal ensemble size and type are, and exactly how the ensemble reduces error and improves generalizability. Furthermore, we plan to utilize the multi-model ensemble to quantify uncertainty in smoke annotations, enabling users like wildfire response teams and environmental agencies to assess the reliability of detections in real time. The application of these ensemble techniques are expected to aid in fire and hazard management by automating the monitoring of smoke in real-time from satellite imagery with smooth and accurate smoke annotations. This will enable improved prediction of wildfire movement and impacts to air quality, ultimately supporting climate resilience and adaptation strategies.

References

[1] Marshall Burke, Anne Driscoll, Sam Heft-Neal, Jiani Xue, Jennifer Burney, and Michael Wara. The changing risk and burden of wildfire in the united states. *Proceedings of the National Academy of Sciences*, 118(2):e2011048118, 2021.

[2] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R Series: A New Generation of Geostationary Environmental Satellites*. Elsevier, 2019.

[3] Donna McNamara, George Stephens, Mark Ruminski, and Tim Kasheta. The hazard mapping system (hms) - noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference on Satellite Meteorology and Oceanography*, 01 2004.

[4] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[5] Thomas G. Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15, 2000.

[6] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[7] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[8] Aurélien Bibaut Cheng Ju and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018. PMID: 31631918.

[9] Giorgio Giacinto and Fabio Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001.

[10] Pavel Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[12] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *CoRR*, abs/1805.10180, 2018.

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

5 Supplementary Material

The code for this work is available at `https://github.com/anonymous-ensemble-smoke/ensemble-AI-smoke-detection/tree/main`. The dataset used will be released in the camera-ready version to preserve anonymity.

An additional example from the test data set is shown in Figure 4, where the individual model output has jagged boundaries and the ensemble outputs smooth over these edges. We see a peak in performance at $N = 8$ in this sample where the $N = 8$ ensemble has the highest IoU score, and the smoothing does not seem to improve in the $N = 12$ ensemble output. This sample supports the proposed idea that ensemble deep learning can smooth over rough edges in semantic segmentation, and warrants further investigation for the optimal ensemble size and type.

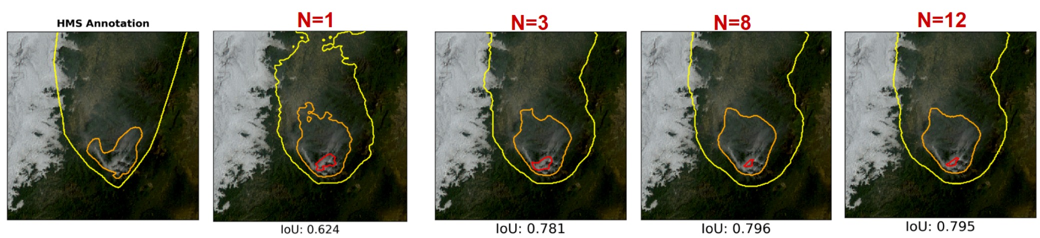

Figure 4: Example of smoke plume detection at (44.24, -122.74) on 2022/09/27 15:30 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. The first panel displays the ground truth annotation; the second panel is the individual model output of DLV3P; the following panels the prediction of an architecture-based ensemble as it increases in size, N.