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

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 With the increasing frequency and severity of wildfires, there is an urgent need
2 for effective and rapid wildfire and smoke detection tools. Recent advancements
3 in computer vision have demonstrated the potential of deep learning models to
4 automatically separate high-resolution images into labeled regions. However,
5 single-model approaches can struggle with generalization and accuracy in diverse
6 conditions. To address these challenges, we propose using an ensemble of deep
7 learning models to produce more accurate annotations of wildfire smoke plumes
8 and their relative density (light, medium, heavy) in Geostationary Operational
9 Environmental Satellite imagery. Our preliminary results indicate that ensemble
10 techniques can improve performance compared to using a single model. This
11 approach aims to provide a more reliable and accurate satellite-based tool for real-
12 time monitoring of smoke, to ultimately aid fire and hazard management efforts
13 and contribute to climate resilience and adaptation strategies.

1 Introduction

15 In the last four decades, wildfire activity has increased drastically in the U.S. In fact, the conditions
16 leading to wildfires have been shown to occur more frequently as a direct result of climate change.
17 Particulate emissions such as smoke and ash are a result of wildfires, and the number of observed
18 smoky days per year has already substantially increased across the continental U.S. Furthermore,
19 future climate predictions reveal that wildfire-related particulate emissions such as smoke and ash
20 could double in fire-prone areas [1]. The human impacts of such smoke exposure include increased
21 morbidity and mortality as well as downstream economic costs [2].

22 Determining causal links between wildfire activity and health impacts in regions distant from the
23 source fire requires accurate large-scale monitoring of wildfire smoke and its movement. An intuitive
24 approach is to utilize satellite-based methods that can monitor evolution of smoke plumes over large
25 areas. However, such methods have yet to provide precise and high frequency information on smoke
26 density.

27 The National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environ-
28 mental Satellites (GOES) provide high spatial and temporal resolution imagery of North America
29 [3]. The NOAA Hazard Mapping System (HMS) Fire and Smoke Product currently relies on human
30 analysts to annotate the presence of smoke over North America using GOES imagery [4]. However,
31 this product is limited by the availability of human analysts and their time. Specifically, annotations
32 are outputted only once to a few times a day and usually have a delay between smoke occurrence
33 and the annotation. Thus, first responders may be delayed when real-time smoke conditions are not

available. To address these limitations, we leverage advancements in deep learning to automate the detection of smoke from GOES imagery, using the existing HMS dataset for training. By automating this task, we can enable more frequent detection of smoke plumes, which will aid in active wildfire monitoring and mitigating air quality impacts.

This proposal focuses on enhancing the capability of deep learning models to detect smoke using multi-model ensemble methods. Ensembles, which combine the predictions of multiple models, have been shown to often perform better than a single model in classification tasks [5]. Particularly, utilizing a diverse set of classifiers in an ensemble is important to achieve the improvement in performance [6]. Furthermore, combining the predictions of multiple independently-trained neural networks can improve generalization and detection accuracy [7–9].

2 Proposed methods

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 seeds to generate different initial random weights. These models were trained independently for 24 hours on 8 Nvidia P100 GPUs using the Adam optimizer and the binary cross entropy loss function.

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. We create an architecture-based ensemble and a initialization-based ensemble. To our knowledge, these ensemble methods have not yet been used for wildfire smoke detection.

2.1 Dataset

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 the validation set and 2022 for the test set, with the remaining years used for training. This ensures the testing and validation data are independent of the training data.

2.2 Metrics

We quantify model performance with Intersection over Union (IoU) score (Equation 1), which measures pixel-wise alignment between the model prediction (y_i^*) and the ground truth (y_i). Validation IoU is used throughout training, and test IoU is used to analyze model performance after training. The IoU metric supports pathways to climate impact since improved IoU score directly translates to more smoke being accurately detected (reducing false detections and increasing true detections) which equips hazard management to appropriately respond to current wildfire and smoke conditions.

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

3 Preliminary results

Table 1 shows the IoU scores on the test set for individual models and ensembles. The ensemble of 8 different architectures outperforms the individual models, with an improvement in all IoU metrics. The ensemble of 8 different initial weights (but the same architecture, PAN) also outperforms the individual models, with a similar improvement in the IoU scores. This improvement is likely due to the different initializations leading to the models searching different parts of the parameter space and

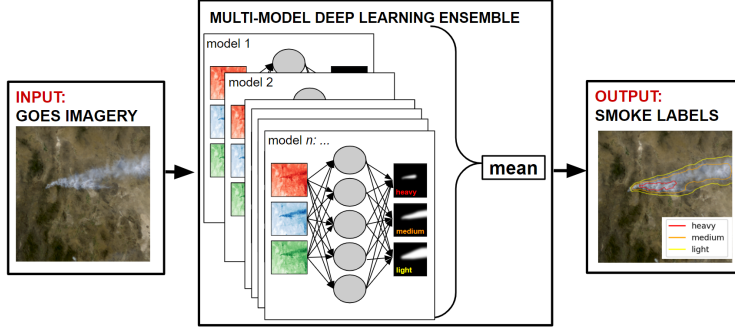


Figure 1: Multi-model ensemble framework. GOES imagery is inputted to N independently-trained models whose output is combined with an un-weighted average to produce the ensemble prediction of pixel-wise smoke labels.

thus finding different minima of the loss function. Future work is necessary to reveal the mechanisms behind the ensemble’s improvement in performance, as well as how to optimally select models that are included in the ensemble.

Figure 2 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: Test IoU results across three classes of smoke density (light, medium, heavy) and over all densities with two single models and two ensemble schemes. N denotes the number of models in the ensemble.

	Heavy	Medium	Light	Overall
Single Model: DLV3P [11]	0.347	0.441	0.666	0.599
Single Model: PAN [12]	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

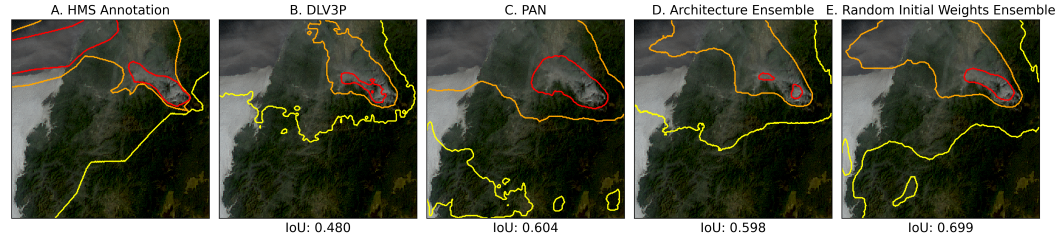


Figure 2: Example of smoke plume detection at (43.37, -123.25) on 2022/10/15 15:50 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. Panel A displays the ground truth annotation; Panels 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 Limitations 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 reveal exactly how the ensemble reduces error and improves generalizability and what the optimal ensemble size and type are. Furthermore, future work will 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.

5 Pathways to climate impact

These ensemble techniques can be used to aid fire and hazard management by automating the monitoring of smoke in real-time from satellite imagery with smooth and accurate smoke annotations. This will improve prediction of wildfire movement and air quality impacts, 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] Wayne E. Cascio. Wildland fire smoke and human health. *The Science of the Total Environment*, 624:586–595, 05 2018. Epub 2017 Dec 27.
- [3] 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.
- [4] 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.
- [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.

6 Supplementary Material

6.1 Data and Code Availability

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.

6.2 Ensemble Size Analysis

Figure 3 shows the IoU performance over all smoke densities as a function of ensemble size, N , for the two ensemble schemes. The ensemble with different initial weights generally improves as models are added to the ensemble. The ensemble of different architectures improves with more models up to 8 models, but then decreased in IoU with more models added to the ensemble. This decrease in performance could be due to the additional architectures not having enough variation in model bias to improve ensemble performance. Future work will aim to clarify exactly how different ensemble sizes behave and reduce error.

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 and how to use the multi-model approach to quantify uncertainty.

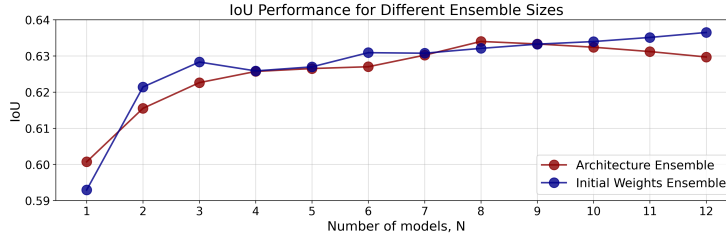


Figure 3: Overall IoU as a function of N for two ensemble design schemes: random initial weights (blue) and architecture-based (red).

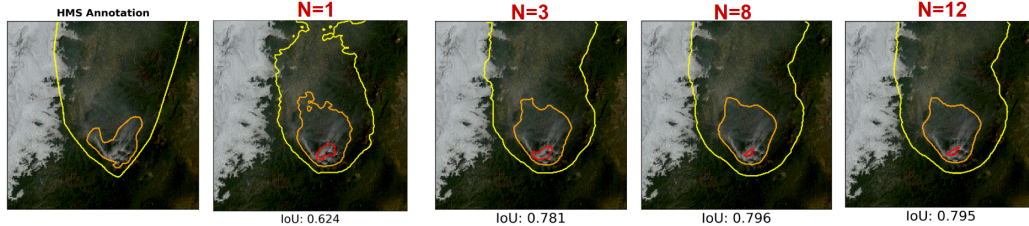


Figure 4: Example of smoke plume detection at (44.24, -122.74) on 2022/09/27 15:30 UTC. Red contours outline the heavy density smoke, orange contours outline the medium density smoke, and yellow contours outline the light density smoke annotations. The first panel displays the ground truth HMS annotation; 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 .