Robust Wildfire Detection with LLM Pseudo-Labeling

Julius Pesonen
Finnish Geospatial Research Institute FGI
julius.pesonen@nls.fi

Abstract

Early detection of wildfires is essential to prevent large-scale fires resulting in extensive environmental and structural damage. Modern deep learning-based computer vision methods enable high-resolution detection of smoke which can be used for early wildfire detection and localisation. Specialised models can even be employed on lightweight drone-carried computers to enable detection in remote areas with low infrastructure. However, such methods suffer from insufficient training data distribution and are thus unable to perform in unseen conditions. This is particularly dangerous for methods intended to recognise emergencies such as wildfires. Multimodal large language models (LLMs) can identify various phenomena in a zero-shot manner offering more robust scaling of the detection domain, but they suffer from computationally heavy inference meaning that specialised models are still required to enable real-time inference with limited computational resources. These models can, however, be trained with LLM pseudo-labelling requiring only unlabelled images and language queries. This builds on previous knowledge of weakly supervised learning for computer vision models through combined vision-language understanding, which has been studied in various contexts. The proposed method could improve the domain adaptability of wildfire detection over previous deep learning-based wildfire smoke segmentation methods. A simple way to formulate the language queries for the pseudo-labelling is through visual question answering (VQA). With feature maps, the language query results can also be transformed into segmentation masks for training models capable of pixel-level detection. LLMs also offer the possibility to evaluate additional features not captured by typical detection or segmentation labels such as rough distance estimates. Using LLMs can thus improve the reliability of smaller detection models in various conditions for which training data collection would otherwise be extremely laborious or even practically impossible.

1 Introduction

Early detection methods of wildfire are essential tools to prevent devastating large-scale forest fires. Drone-based detection offers avenues for quick deployment of surveys on large areas with low infrastructure. In remote areas, however, the drones are limited to on-board computing for detection due to the lack of high-bandwidth mobile networks.

Specialised models are required to enable real-time early detection with on-board resources. For learning based approaches a major limiting factor, however, is the narrow distribution of publicly available data, especially labelled data. Recently, models combining vision and language modalities have risen as tools for automating image label generation and LLMs have been increasingly incorporating the visual input modalities. Due to their extensive capabilities in language expressions, they also offer improved capabilities in image understanding and image-language tasks such as VQA.

Key Benefits of LLM Pseudo-supervision

For computer vision:

- No manual labelling
- Easier data scaling
- Improved explainability

For language models:

- Evaluation based on teaching ability
- Improved explainability

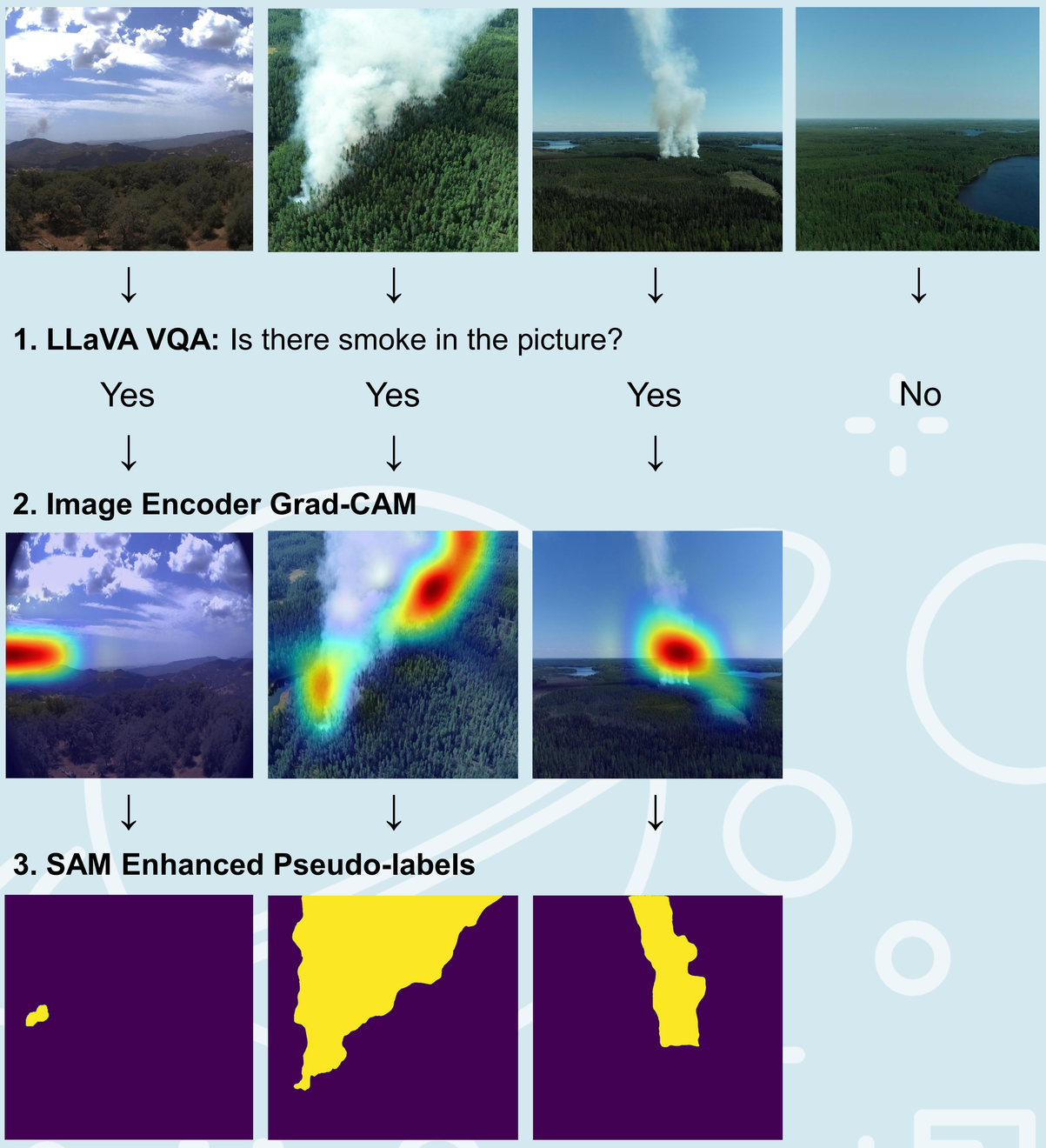

***Figure 1**: The full pseudo-label generation process from unlabelled images to dense smoke masks for inference model supervision.*

2 Methods

The proposed method consists of the pseudo-label generation process, visualised in **Figure 1** and described step by step in **Table 1**, and the inference model training.

Step	Description
1. VQA with LLaVA-v1.5-7B [1]	The model is provided images and asked the question: *"Is there smoke in the picture?"* The answer is used as a binary pseudo-label. The next steps are only applied for positive samples.
2. Grad-CAM of the image Encoder [2]	The class activation maps of the image encoder are used to determine the location of the detected smoke in each image and initial pseudo-masks are generated.
3. SAM enhancement [3,4]	The pseudo-masks are enhanced using the Segment Anything model.

***Table 1.** The proposed pseudo-label generation steps.*

Initial tests on the ability of the LLM to detect smoke and the demonstrative visualisations were made using smoke images by AI For Mankind and HPWREN [5] and our own drone collected data introduced by Raita-Hakola et al. [6]. In addition, two more VQA queries were used to estimate the model's capability to predict the distance to the detected smoke in open text form (**VQA 2.1**) or in meters (**VQA 2.2**).

3 Results

The smoke detection accuracy is shown in **Table 2** and sample images are displayed in **Figure 1**. For **VQA 2.1** the answers presented slight contradictions, such as "The smoke is relatively close to the image-capturing location, as it is seen in the background of the image." for the leftmost sample of **Figure 1**. For the examples in **Figure 1** the model estimated the metric distances to the smoke as 1 000, 100, 1 000 and 1 000 meters. Interestingly, in the smokeless images the Grad-CAM also presents hallucinations like the distance VQAs.

Model	Precision	Recall	F1 score
LLaVA-v1.5-7B	0.862	0.882l	0.926

***Table 2:** LLaVA-v1.5-7B smoke detection results. The metrics were computed on a test set of 120 images, 80 with smoke and 40 without.*

4 Conclusion

As shown by the results in **Table 2** and the small number of samples in **Figure 1** the multimodal LLM can distinguish images with smoke from an unseen distribution of images. The proposed pseudo-label generation method shows a possible way of distilling this knowledge into specialised segmentation models for practical inference.

The **VQA 2.1** and **2.2** also demonstrated how additional information could be extracted from the images using LLMs. However, the presented contradictions and hallucinations show the importance of starting the process using a simple binary question providing the answer weather the desired object appears in the image at all.

References

[1] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296– 26306, 2024.
[2] B. Zhou, A. Khosla, `A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization, "2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929, 2015.
[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.
[4] T. Chen, Z. Mai, R. Li, and W.-l. Chao, "Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation," arXiv preprint arXiv:2305.05803, 2023.
[5] AI For Mankind, "Open wildfire smoke datasets." https://github.com/aiformankind/wildfire-smoke-dataset, 2020.
[6] A.-M. Raita-Hakola, S. Rahkonen, J. Suomalainen, L. Markelin, R. Oliveira, T. Hakala, N. Koivumäki, E. Honkavaara, and I. Pölönen, "Combining yolo v5 and transfer learning for smoke-based wildfire detection in boreal forests," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 48, pp. 1771–1778, 2023.

Acknowledgements

This research was funded by the Academy of Finland within project Fireman (decision no. 346710, 348009). The FireMan project is funded under the EU's Recovery and Resilience Facility that promotes the green and digital transitions through research. This study has been performed with affiliation to the Academy of Finland Flagship Forest–Human–Machine Interplay—Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences (UNITE) (decision no. 357908).