Proposal Report

Sepehr Heydarian, Archer Liu, Elshaday Yoseph, Tien Nguyen

Table of contents

1 Abstract **1**

2 Introduction **1**

3 Proposed Pipeline **1**
3.1 High-Level Overview of the Proposed Data Pipeline 1
3.2 The Data 3

4 Timeline **3**

5 References **3**

1 Abstract

Hello

2 Introduction

3 Proposed Pipeline

3.1 High-Level Overview of the Proposed Data Pipeline

With a large amount of anticipated unlabelled imags coming in, a major challenge is how to keep the model current and effective as new data arrives. Manual labeling and retraining will soon no longer be sustainable, especially under the time pressures involved in wildfire detection.

To solve this, we propose an intelligent and iterative data pipeline. This solution combines automation, human oversight, and model optimization to support continuous improvements. Below is a high-level summary of the process:

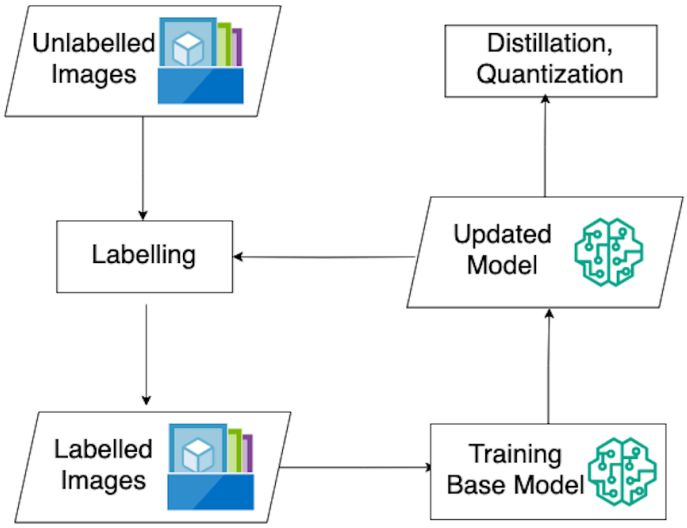

Figure 1: High-level overview of the data pipeline

1. **Unlabelled Image Ingestion**
   The pipeline begins by collecting raw, unlabelled images from the client's data sources.
2. **Automated Pre-Labelling**
   One or more AI models are used to perform initial labeling. They identifies objects such as fire, smoke, or vehicles. This step helps reduce the time and effort required for manual annotation.
3. **Human-in-the-Loop Review**
   Human experts then review and correct the AI-generated labels. This step ensures high accuracy while avoiding the burden of fully manual labeling.
4. **Model Retraining with Verified Labels**
   The corrected labels are used to retrain or fine-tune the main wildfire detection model. This helps the model learn from new data and maintain strong performance over time.
5. **Model Optimization for Deployment**
   Optimization techniques such as **distillation** and **quantization** are applied. These help reduce the size and complexity of the model, making it more efficient for deployment on edge devices.

Keypoint: This pipeline is designed for continuous learning. Each iteration allows the system to improve, leading to better detection accuracy and faster response in real-world situations.

Further details on the technical components and implementation will be provided in the following sections.

3.2 The Data

The proposed pipeline is designed to process image data annotated with bounding boxes in text format. The client currently maintains a repository of over 2 million unlabelled images, with approximately 500 new unlabelled images expected to be added each month from various sources. While the full dataset is stored on Google Cloud Storage (GCS), for the purpose of the prototype, we assume the data is stored and accessed locally to simplify development and testing.

To support early experimentation, our team has been provided with a mixed dataset consisting of both labelled and unlabelled images. The object detection model will focus on five key classes: Fire, Smoke, Lightning, Vehicle, and Person.

4 Timeline

5 References