Multi Sensor Urban Air Quality Prediction

Aditya, 2021CE10494

1 Abstract

Urban air pollution poses significant threats to public health and quality of life. Traditional air quality forecasting systems rely on a single data modality (e.g., weather data or isolated pollutant sensors), potentially overlooking rich spatial and contextual patterns. This project proposes a multi-sensor deep learning framework that integrates diverse data sources—including ground-level sensor readings, satellite imagery, and weather forecasts—to predict urban air quality with higher accuracy and robustness. By leveraging graph-based architectures and convolutional neural networks, the aim is to capture spatiotemporal dependencies across a city-wide sensor network.

2 Introduction

Air pollution has been recognized as a serious global issue, contributing to millions of premature deaths annually and imposing substantial economic costs [1]. Predictive modeling of urban air quality enables early warnings and proactive interventions, such as traffic regulation or industrial emission control. Existing approaches often fail to capture the multi-faceted nature of urban environments, where topography, land-use patterns, and localized pollution events can vary significantly within a short distance. Deep learning architectures have demonstrated success in learning complex dependencies, motivating the exploration of multi-sensor integration to enhance predictive accuracy [2].

3 Problem Statement and Key Challenges

Problem Statement: Predicting daily and hourly air quality indices (e.g., PM2.5, NOx, or AQI) across a city by fusing heterogeneous data sources: ground-based sensor readings, satellite imagery, traffic flows, and meteorological forecasts.
Key Challenges:

- **Data Heterogeneity:** Combining data from sensors, satellite imagery, and weather data requires careful preprocessing and alignment.
- **Spatiotemporal Correlations:** Pollution levels exhibit both geographic (across sensor locations) and temporal dependencies (hourly/daily).
- **Missing and Noisy Data:** Sensor malfunctions, cloud cover in satellite imagery, and reporting delays can degrade performance.
- **Model Complexity vs. Interpretability:** Deep models can capture complex patterns but often act as black boxes, complicating trust and regulatory acceptance.

4 Description of the Datasets

The project will integrate the following publicly available datasets:

- **OpenAQ** [3]: Real-time and historical air quality measurements (PM2.5, NO2, O3, etc.) from multiple urban stations worldwide.

- **NOAA Weather Data**: Historical weather features such as temperature, humidity, wind speed, and precipitation.
- **Satellite Imagery (Sentinel/Landsat)**: Remote sensing data (land-use, vegetation indices, aerosol optical depth) accessed via Google Earth Engine.
- **Traffic Flow Data**: Open data portals (e.g., city-level) providing hourly or daily traffic intensities for urban regions.

Datasets will be synchronized by timestamp and geolocation. Sensor stations serve as nodes in a graph, while imagery is sampled in local patches surrounding each station.

5 Proposed Methodology

Data Preprocessing and Fusion:

- *Spatial Alignment:* Map satellite pixel data to sensor coordinates, possibly using a fixed radius around each sensor location.
- *Time Alignment:* Aggregate measurements into consistent time windows (e.g., hourly).

Model Architecture:

- *Graph Neural Network (GNN)*: Nodes represent sensor locations, edges model proximity or shared traffic routes. This layer captures spatial correlations.
- *CNN for Satellite Images*: A small convolutional network extracts features from local patches of satellite imagery, encoding land-use or aerosol patterns.
- *Temporal Component (LSTM/Transformer)*: Learns temporal trends from historical sequences of sensor readings and weather data.

Training and Evaluation:

- *Loss Function*: Mean squared error or mean absolute error for pollutant concentration predictions.
- *Metrics*: Root mean square error (RMSE) and correlation coefficients for air quality indices.
- *Experiments*: Comparative analysis with baselines (e.g., simple linear regression, single-modality CNN).

Expected Outcomes: A unified model that outperforms single-source methods, offering more accurate short-term forecasts. Insights from attention mechanisms or GNN weights might highlight the most critical factors (e.g., traffic density vs. weather patterns).

References

References

[1] World Health Organization. "Ambient air pollution: A global assessment of exposure and burden of disease," 2016.

[2] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," *KDD*, 2013.

[3] OpenAQ: `https://openaq.org/`, accessed 2025.