Project Specification: Incorporating deep learning models into the UrbanAir system

Raihanah Lukman

October 2023

1 Background

Measuring air quality is crucial as it directly impacts us with poor air quality leading to an estimated 9000+ Londoners' premature deaths every year and even contributing to climate change. Monitoring air quality helps policymakers develop effective strategies, implement regulations and reduce pollution by helping them make more informed decisions. While London has made strides in improving air quality, air pollution levels persist in certain areas, posing a severe health risk to residents.

The UrbanAir system currently uses a Gaussian Process (GP) model which is a probabilistic model for machine learning and statistics that represents a distribution over functions. It's defined by a kernel function that encodes the relationships between input points and their corresponding function values. More specifically, the UrbanAir system incorporates Multi-Resolution Deep Gaussian Processes (MR-DGP) and Sparse Variational Gaussian Process (SVGP).

Below shows a comparison of Multi-Resolution Gaussian Process Regression Networks (MR-GPRN) and MR-DGP models [1].

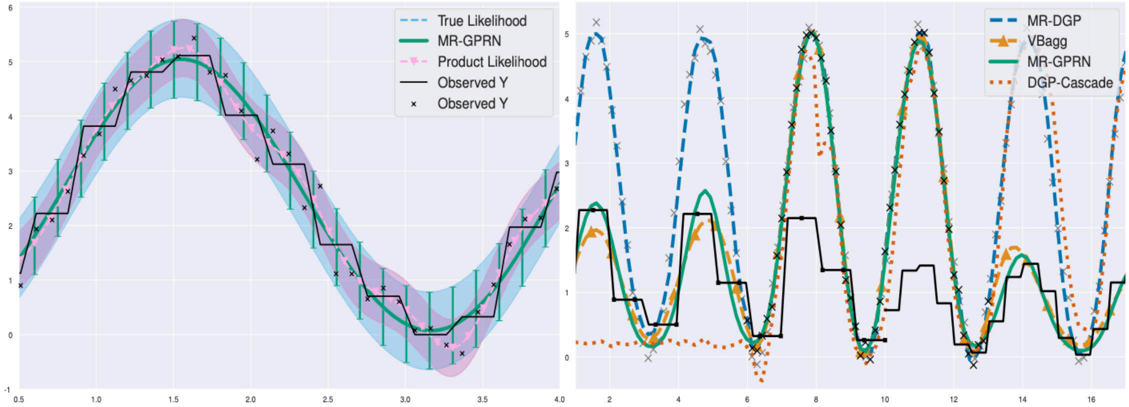


Figure 1: **Left**: MR-GPRN recovers the true predictive variance whereas assuming a product likelihood assumption leads to posterior contraction. **Right**: MR-DGP recovers the true predictive mean under a multi-resolution setting with scaling biases. Both VBAGG-NORMAL and MR-GPRN fail as they propagate the bias. Black crosses and lines denote observed values. Grey crosses denote observations removed for testing.

2 Problem Statement

GP models offer the ability to model intricate relationships by incorporating expert knowledge via the choice of kernels but face limitations when working with high dimensional problems due to it being computationally expensive [2] which is the case in the UrbanAir system which leverages city-wide air quality sensors, measuring spatiotemporal features to develop machine learning algorithms and data science platforms aimed at understanding and enhancing air quality in London. This project aims to integrate Deep Neural Networks (DNNs) such as Convolution Neural Network - Long Short Term Memory (CNN-LSTM) which are widely employed due to their ability to scale efficiently and excel in point predictions compared to GP models into the UrbanAir system and evaluate the potential benefits and drawbacks of using DNNs compared to the GP model.

3 Objectives

The objective of this project is to:

1. Build a further understanding on DNNs
    - Performing literature reviews on similar projects
    - Researching Neural Networks (NNs) & DNNs, specifically Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) for the sequential data and CNNs for the spacial data
2. Incorporate Deep Neural Networks (DNNs) into the UrbanAir system, leveraging their widespread use for efficient scalability and superior performance in making specific predictions when compared to GP models.
    - Selecting a DNN structure that can handle spacial-temporal data (CNN-LSTMs)
    - Building a model that can predict air pollution levels using time series data using RNN and LSTMs
    - Building a model that can predict air pollution levels using time series & location data using CNN-LSTM
    - Evaluating the accuracy & performance of the DNN in predicting air pollution levels
3. Assess the advantages and disadvantages of employing DNNs in contrast to the GP model within the UrbanAir system.
    - Build understanding on the current GP model within the UrbanAir system
    - Identify weaknesses and strengths of GP model
    - Identify weaknesses and strengths of DNN model
    - Compare and contrast both models and identify advantage and disadvantage of both models

4 Methods & Methodology

Overall, a mixture of planned and agile methodology will be adopted for different phases of the project which are research, development & testing, and evaluation. The aim would be to create an iterative process that produces neural networks of satisfactory accuracy and functionality.

4.1 Phase 1: Research

Before any development can occur, a thorough understanding of DNNs for spacialtemoporal data and the project must be demonstrated. Any research materials will be checked if it is relevant to the current project first before spending time on it and any parts that I do not understand will be taken note of to be communicated to my supervisors.

1. Communication with my supervisors Theo and Sueda will be done via Slack and weekly meetings have been set with both of them on 11am Tuesdays and 10am Fridays to ensure everything is on track and any concerns I have may be addressed.

2. Doing literature review on similar research projects to familiarise myself with the process and topics surrounding my project

3. Understanding the dataset and data sources

4. Research on RNNs, CNNs and LSTMs to select the most suitable structure to use for the project

4.2 Phase 2: Development & Testing

The development phase will be run with the goal of producing a working DNN that can accurately predict air quality. The process will be run iteratively with improvements being introduced with each iteration and accuracy testing being done at the end of each iteration. If I find myself stuck at any point, I will reach out to my supervisors for help.

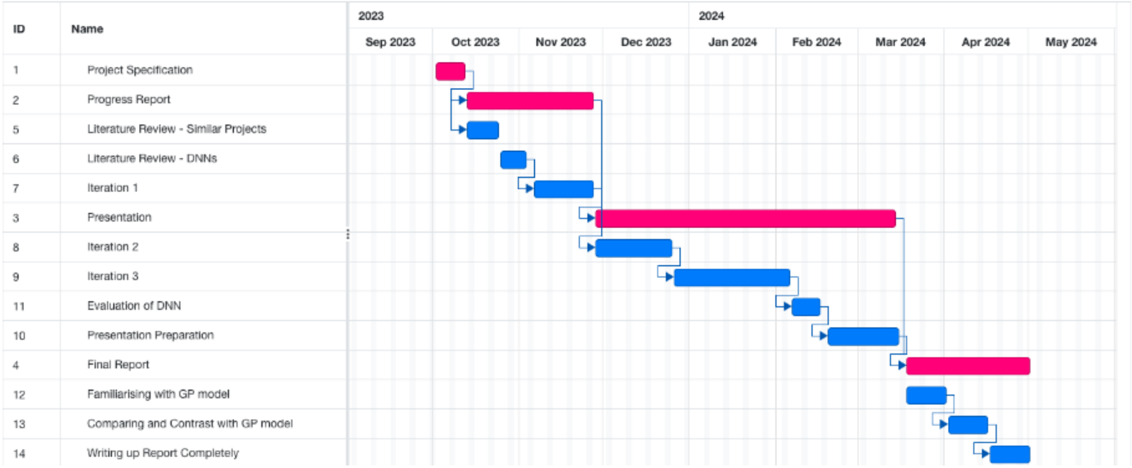
1. Iteration 1: Implement simple Neural Networks on toy settings to familiarise myself with the process
2. Iteration 2: Implementing RNN to predict air quality using some features (a simplified version of the actual problem, such as only predicting one type of pollutant or using less features than the actual problem)
3. Iteration 3: Implementing a LSTM that can predict air quality on a simplified dataset using time series data
4. Iteration 4: Implementing a CNN-LSTM that can predict air quality on a simplified dataset using spacialtemporal data
5. Iteration 5: Implementing a CNN-LSTM that can predict air quality as intended by the actual problem

4.3 Phase 3: Evaluation

Once a satisfactory DNN model has been produced, we would then want to compare it with the current GP model to see the advantages and disadvantages of each model. This phase will be done and executed only if there is sufficient time to do so.

1. Familiarising myself with the performance, advantages and disadvantages of the GP model
2. Familiarising myself with the performance, advantages and disadvantages of the DNN model
3. Comparing both models

5 Timetable



6 Resources & Risk

6.1 Resources

The project is an extension of the London Air Quality Project under the Alan Turing Institute and will be using data from the London Air Quality Network and Copernicus satellite air quality forecasts. The code will be written in Python using Jupyter, tools such as TensorFlow will also be utilised. The project will be run on personal, non-DCS, machines for the most part but the DCS GPU machines may be used periodically to reduce computational time. GitHub will also be used as a resource for any files used in the UrbanAir system. Literature includes books on Neural Networks such as Neural Netoworks for Pattern Recognition, similar projects publications as well. Any additional resources required from the Alan Turing Institute will be relayed to my supervisor.

6.2 Risk

Risk: Getting stuck at a certain part of the project and it taking longer than expected due to illness, underestimating the problem, other priorities such as unpredictable events or other academic work taking precedence.

- If I am stuck due to lack of knowledge/skill I will reach out to my supervisors for help
- Other issues such as illness or unpredictable events will be mitigated as it occurs.
- Proper planning and balance of other modules will be needed to avoid other academic work taking over the time planned for the project

Risk: The DNN structure chosen fails to give satisfying results

- Doing proper research before choosing the structure to minimise risk of choosing the wrong one
- Doing periodic checks during development to ensure that the DNN is progressing as hoped and has potential to give satisfying results
- Raising any concerns to supervisors for advice immediately

Risk: As the model becomes increasingly complex, it may take a longer time to compute and overload my personal machine

- Use the DCS GPU machines if this happens

Risk: Spatial and temporal correlation varies across space and time due to factors like non-stationarity, non-separability

- Utilise plots such as Autocorrelation/PartialAutocorrelation (ACF/pACF) plots, (semi-)variograms etc

7 Legal, Social, Ethical and Professional Issues & Considerations

This project does not require any human data or participation so there is no concern in that aspect. However I will be using data from air quality sensors placed around London that is open to the public.

We also need to take into account the Uncertainty Quantification (UQ) especially since there is a lack of ground truth data and interpretability of how the model comes up with the result. UQ is essential in facilitating risk assessment by quantifying potential harms and uncertainties associated with technical, financial, and environmental aspects. Hence, I will make sure to transparently convey uncertainties to stakeholders by including a disclaimer in this project, allowing them to make well-informed decisions about their involvement in the project.

References

[1] O Hamelijnck. “Multi-resolution Multi-task Gaussian Processes”. In: (2019).

[2] Oscar Knagg. “An intuitive guide to Gaussian processes”. In: (2019).