Forecasting Long-Term Urban Air Quality Index using Multi-Model

Khanh-Linh Vo[1,*,†], Gia-Nghi Phuc-Nguyen[1,*,†], Tuong-Nghiem Diep[1,*,†] and Nhat-Hao Pham[1,*,†]

[1]*VNUHCM - University of Science, Vietnam*

Abstract

Forecasting and assessing pollutants values accurately is always an attractive topic in the research community during the last decade. It will help to provide a good approach to problems related to health effects associated with current air quality conditions. In this paper, we will provide 2 approaches utilizing machine learning to solve 2 subtasks. In subtask 1, we use multi-models to forecast AQI value and in subtask 2, we use VGG16 model which was customized to predict AQI level using only pictures. We evaluated the performance of two models using the dataset from sensors stations installed across Dalat City, Vietnam. The experiment results show that our proposed models yield high performance in terms of MAE, MSE, and F1-score.

1. Introduction

In Urban Air: Urban Life and Air Pollution task[1], MediaEval2022, we participated in both subtasks: Forecast pollutant values then predict Air Quality Index (AQI) and predict AQI using pictures only then forecast AQI in the future. The target of both subtasks is to forecast AQI in the mid-term and long-term using data from the locally available device. We propose using multi-models in task 1: Fourier model was used to forecast some weather components, after that, we use these features to predict the index of air pollutants to predict AQI for each pollutant. In subtask 2, we use VGG16 and a concatenate layer at the backbone for solving this task which only pictures for use, we use the nearest camera provided by CCTV of sensors to predict the AQI level at these local sensors. In this model, we use pictures from the same time of 7 days ago to predict.

2. Approach

2.1. Subtask 1

After some visualization stages, there are several issues with provided data: untrusted data, missing data, and duplicated data. With untrusted data and missing data, we just simply remove all of them from our dataset. With some sensors (such as sensor 1), the amount of data missing was not too much to affect the result but with others, there are approximately 80% amount of data was missing so we calculated a ”base” value and applied it to our missing-timestamp data.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

†These authors contributed equally.

21280070@student.hcmus.edu.vn (K. Vo); 21280035@student.hcmus.edu.vn (G. Phuc-Nguyen); 21125155@student.hcmus.edu.vn (T. Diep); 21280066@student.hcmus.edu.vn (N. Pham)

Because of missing data and the time of making the prediction was the time of intersection between the rainy and dry seasons can cause rapid weather condition changes so we decided to use a window time of 1 week.

Our used method contains 3 steps: Forecast weather components, predict pollutants' value, and predict AQI based on the value of pollutants.

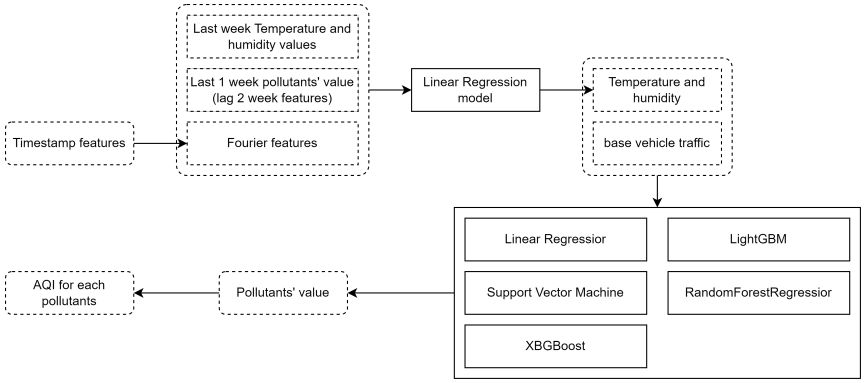

Figure 1: Subtask 1 pipeline

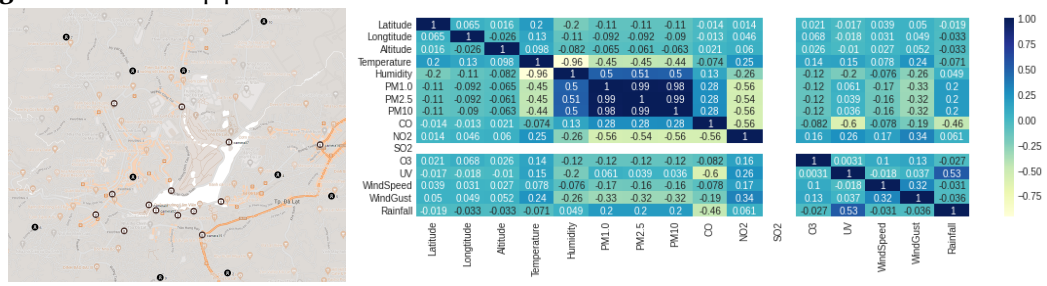

Figure 2: (left) location of sensors and cameras, (right) correlation map of features

Because temperature is a regular feature, we used Fourier-transform to forecast the temperature values for the forecasting week, and after that used it to predict the humidity feature because of a strong correlation (Figure 2 -right) between them.

In the second stage, we used features forecasted and predicted in the last step to predict the value of pollutants. Because several models cause negative predictions, after all, we choose XBGBoost[2] to predict the value of pollutants.

In the last stage, we predict AQI for each pollutant based on [3]. Some pollutants required an 8-hour value to calculate AQI, but we mimic it like 1-hour as the guidance of the task organizers.

According to the requirements of the organizers, the pollutant indicators are calculated through AQI according to EPA standards. However, the EPA does not determine the AQI value for O3_1H when the index is less than 0.124, so the reference becomes unreasonable. So we propose a slight modification to the calculation of the AQI value for O3_1H with a value less than 0.124 according to the following rule: AQI 50, 100 corresponding to 0.08 and 0.124 O3_1H respectively.

2.2. Subtask 2

We approach this subtask using the image only. Firstly, we think about solving the problem with features extracted by hand including a number of vehicles and pedestrians, but we realize that this work does not ensure whether these features are good and we believe inexact features

will get us the wrong way to approach this task and make the model potentially lead to inaccurate predictions. Therefore, we decide to solve this task using Multiple inputs and Multiple models such as VGG16 to predict each pollutant level because we would like features to be extracted by robust models through learning from data.

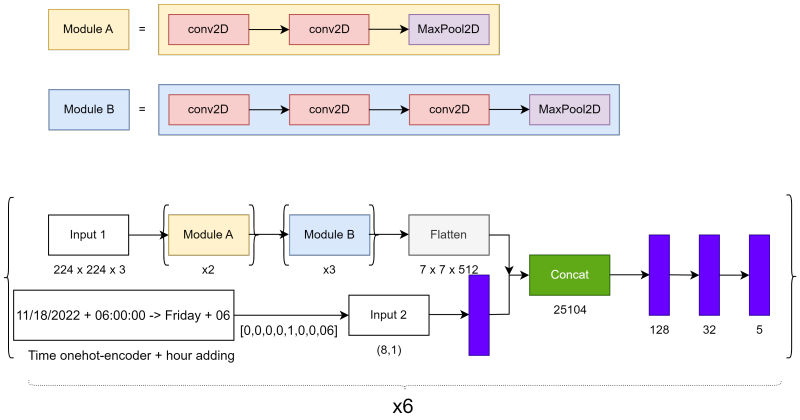

Figure 3: Subtask 2 model. In the VGG16 model, there are two main modules that make VGG16 robust, Module A and Module B

In the model, we use VGG16[4] concatenated[5] with a time vector at the flattened layer of VGG. With time features, we one-hot-encoder[6] the day in a week and then combine it with the nearly hour time, see in figure 3. We generate a time series feature for this model.

For each AQI pollutant level, we use this model architecture and train on a dataset including the AQI level of each pollutant of the sensor and CCTV nearest sensor (Figure 2 -left). We generate 6 models in the same architecture for predicting AQI pollutants - PM2.5, PM10, CO, O3, SO2, and NO2.

This model was based on Classification problems, we classify 5 levels of AQI pollutants - PM2.5, PM10, CO, O3, SO2, and NO2. Before adding the vector time feature, we tested it on only VGG16. The loss and accuracy of this baseline model are not good. After adding the vector time feature, this new model predicts better than.

We train it on 20 epochs and we realize, the original VGG16 does not increase accuracy or loss, but our new model can approach 0.9 while the epoch comes over 100. The optimizer is Adam, we evaluate in sparse categorical cross-entropy loss. The activation function is Relu.

With each sensor, we use images from its nearest camera and these images we use are from 7 days ago or days near the time which we need to predict.

3. Result

Pollutant	MSE	F1-score
PM2.5	388.17	0.69
PM10	91.38	0.97
CO	5.87	1
NO2	0.33	1
SO2	1	1
O3	108.0	0.82

Table 1
MSE (for values) and F1-score (for levels) of each pollutant in subtask 1

According to table 1, chemical pollutants have higher prediction performance than particulate matter values. This can be explained by some problems related to missing data and the lack of some necessary features (such as traffic being constantly updated because the request for the subtask is not allowed). Moreover, values of O_2, CO_2, and SO_2 in Dalat as we know when visualizing them are always within a fixed range of AQI level such as good level, so in terms of ideas, we do not have to predict it.

In subtask 2, We evaluated the VGG16 on image features only and saw that the performance is not good. We thought that because image features are not enough and the time we need to predict AQI level is an important factor. Therefore, we tried to add additional encoded timestamp features and chose a small but comprehensive model, VGG16, that can capture useful features. Adding the time feature, in combination with our experimental evidence, has shown that our hyperthesis is correct. The addition of these timestamp features has led to an increase in accuracy and F1 score.

Model	Accuracy	Loss	F1	Precision	Recall
VGG16	0.71	0.66	0.54	0.41	0.88
VGG16+ Time	0.80	0.47	0.56	0.44	1.0

Table 2
Compare baseline and adding time feature in subtask 2

Instead of using VGG16, with our conclusion, it can be replaced by more powerful models such as ResNet, DenseNet, ..., however the results will not be much better, the important thing here is the hand-features added.

4. Conclusion and outlook

In this competition, we used multi-models to forecast AQI for each pollutant. Because of several reasons, our performance was not as good as we had hoped but we have gained a lot of experience by participating in this task.

When building the correlation hypothesis between traffic, weather, and air pollution, both people's experience and the model's knowledge played important roles in the process. The difference between them is the source of the information being used and the level of subjectivity involved. People's experience is based on their observations, learning and sometimes, ideas come from hunch. In fact, we are Vietnamese and have been to Dalat city many times, so that, we know which human activities often cause air pollution, which areas are crowded, as well as geographical, weather, and seasonal factors, but the prediction can be influenced by our own preconceptions. While a model's knowledge is based on data and information that has been input into the model. When the data was collected incorrectly, no matter how good the model is, it cannot make the right prediction. However, when the data is "acceptable", the model can identify patterns and trends that may not be immediately apparent to the human observer, just based on data and following logical rules and algorithms. The similarity between people's experience and model's knowledge is that both can be used to identify patterns and trends, test and validate the hypothesis. Therefore, the two are often used together to aid in the hypothesis-building process.

In the future, we think that we can have some improvements to our methods and maybe can get higher performance. For subtask 1, we can combine it with real traffic features (not now because we still cannot get the exact traffic from CCTV) to calculate the traffic flow over the city.

References

[1] M.-S. D. T.-H. D. T.-L. N.-T. T.-B. N. D.-T. Dang-Nguyen, Overview of mediaeval 2022 urban air: Urban life and air pollution (2022).

[2] T. C. C. Guestrin, Xgboost: A scalable tree boosting system (2016).

[3] EPA, Technical assistance document for the reporting of daily air quality – the air quality index (aqi) (2018).

[4] K. S. A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014).

[5] H. Simon, A multilayer perceptron is a class of feedforward artificial neural network (1994).

[6] B. Jason, Why one-hot encode data in machine learning (2017).