Report for EMAI Challenge

Team: A-P10005

January 14, 2022

1 Introduction

To predict the hourly cooling load out of the four given natural parameters namely temperature, humidity, UV index, and rainfall with 15 min intervals, we design a sequence-to-sequence model based on self-attention mechanism. We also conduct a series of data analyses on the provided dataset which gives us better insights on the data behavior on a statistical view. To fully utilize the given data, we perform data restoration on both the cooling load and other parameters. Finally, our model achieves rmse below 200 and l1loss below 180 on randomly selected test set.

2 Data Processing

The behavior of cooling load has a close relationship not only with the natural parameters but also with the working hour of humans. To get a better understanding of this relationship, we conduct a comprehensive data analysis before we get down to design the model. There are also a lot of outliers, missing data in the dataset, which exists not only on natural parameters and cooling load itself. We designed two approaches to tackle these problems separately.

2.1 Data Analysis

The cooling load actually reflects the willingness that people in this building of turning the AC on. Thus, both the climate and human activity can have an impact on the cooling load. For example, when the temperature is high, we can see a higher cooling load vise versa, and when people come for work, we can see a higher cooling load. Thus, the cooling load not only has a seasonal pattern but also shows significant periodicity on day and week, as shown in fig.1.

2.2 Data Imputation and Retrieving Public Data

There are a large number of missing data on the cooling load. Some of them are the result of incorrect sensor readings which are rare and sparse. But there also exist large chunks of missing data, eg, 2000+ missing data points on the July and October of 2020, which is about 2 weeks of missing. For the former case, we use a simple interpolation method to restore the missing data points. But for the latter case, we have no choice but to discard them.

Missing data also exists in the natural parameters. Such missing data makes the precious cooling load information unable to serve as the training data. Instead of using the brute interpolation method, we choose to pull the public climate data from HK Observatory API, since the climate should be identical across Hong Kong.

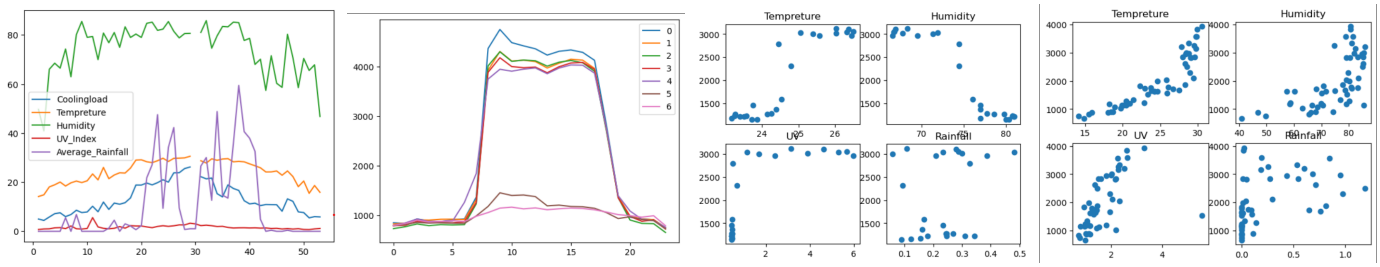

Figure 1: The right-most figure shows the waveform of five parameters. The second shows the average waveform of each day of the week, which is also the prototype of each day. The following two figure shows the relationship between four climate parameters and the cooling load on the scale of day and year.

DayofYear	Date	DayofWeek	Outlier type
266 269	2020/9/22 25	Tue Fri	Cooling load is about 4k while the next week is about 7k
144, 145, 152	2020/5/23, 24, 31	Sat, Sun, Sun	Weekends which have a constant cooling load of 2250
165	2021/6/14	xxx	Public holiday but waveform shows it's a work day
182	2021/7/1	xxx	Public holiday but waveform shows it's a work day

Table 1: Outlier in Dataset

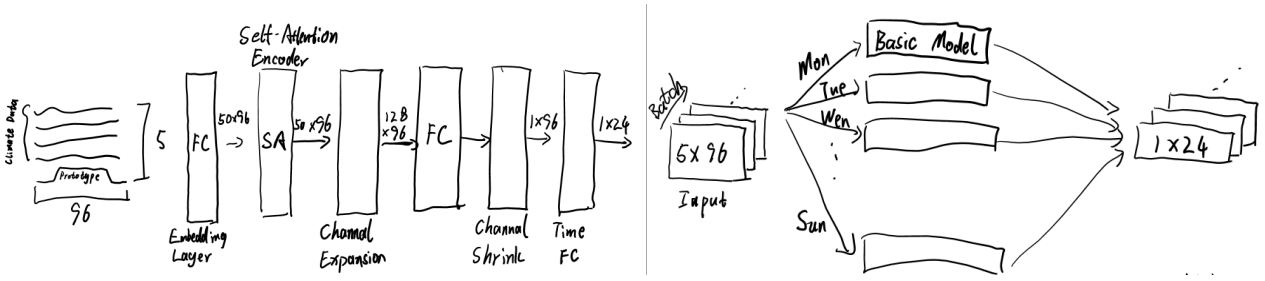

Figure 2: The figure on the left shows the design of our basic model. The left figure shows the

2.3 Cross Validation

We perform cross-validation on the whole dataset to find out if each part of the dataset shares the same attributes. We split the dataset into 6 parts and each time take one part as the testing set and the other five parts as the training set. We find several outliers in the dataset, shown in table.1.

3 Make Use of Timestamp

Both the natural factor and mankind factor can affect the cooling load. The natural factors are given by the natural parameters, while the mankind factor is hidden in the timestamp.

By visualizing the average waveform of the Monday, Tuesday, etc as shown in [fig.1], we find that there are the difference among them, especially for the weekend. To tackle this problem, we extract the dayofyear and dayofweek from timestamp to determine if a particular day is a workday and check if that day is a public holiday.

To differentiate these days, we introduce a new input channel called the 'Prototype'. The prototype is basically the average hourly cooling load of a certain day of the week, e.g. the prototype for Monday is the average of all the Mondays in the dataset. The prototype not only provides labels for different days but also introduces detailed prior knowledge that what the waveform of a certain day should be like. Specially, we set the prototype of public holiday as the prototype for Sunday because legally Sundays are also public holidays and we expect the waveform of them to be identical.

4 Model Design

Since the cooling load should be identical for days with the identical day of the year because the natural parameters should be similar. For example, the cooling load for 2021/9/1 should be similar to the cooling load around the beginning of September of 2020. Based on this assumption, we choose the self-attention layers as our model encoder. And we use several fully connected layers as the decoder.

We notice that sharing the same decoder makes it hard to predict well both on workday and holiday. Thus we trained two networks to predict them separately. Furthermore, we find it performs better when training seven networks which each of them learns the pattern of one single day of the week.