به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین چهارم

نام و نام خانوادگی	امیرحسین پورداود
شماره دانشجویی	۸۱۰۱۰۱۱۲۰
تاریخ ارسال گزارش	۱۴۰۱.۱۰.۵

فهرست

شکل‌ها

جدول‌ها

پاسخ ۱. تخمین آلودگی هوا

۱-۱. سوالات تشریحی

در این قسمت متد های زیر را بصورت مختصر شرح میدهیم:

۱-۱-۱. Linear interpolation method[1]

در linear interpolation، دو نقطه داده با یک خط مستقیم به یکدیگر متصل میشوند که بنابراین تا بع درونیابی بصورت زیر خواهد بود:

$$f_1(x) = b_0 + b_1(x - x_0)$$

که در آن x یک متغیر مستقل هست، x_i ; i = 0, 1, 2, … ها یک مقدار مشخص از متغیر مستقل است و b_i ها یک ضریب نامشخص است. و از فرمول بالا داریم:

$$b_0 = f(x_0)$$

و

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

که در آن ها $f = f_1$ میباشد.

در صورت وجود ۳ نقطه، درونیابی بوسیله چندجمله ای مربع (درجه ۲ بدست می آید که رابطه آن بصورت زیر بدست می‌آید:

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

ضرایب b_0 و b_1 طبق فرمول ذکر شده در بالا بدست می‌آید که $f = f_2$ است و b_2 برابر است با:

[1] Norazian, Mohamed Noor et al. “Estimation of missing values in air pollution data using single imputation techniques.” (2008).

$$b_2 = \frac{\frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2-x_0}$$

در صورت وجود ۳ نقطه، درونیابی بوسیله چندجمله ای مکعب (درجه ۳) بدست می آید که رابطه آن بصورت زیر بدست می‌آید:

$$f_3(x) = b_0 + b_1(x-x_0) + b_2(x-x_0)(x-x_1) + b_3(x-x_0)(x-x_1)(x-x_2)$$

ضرایب b_0 و b_1 و b_2 طبق فرمول ذکر شده در بالا بدست می‌آید که $f = f_3$ است و b_3 برابر است با:

$$b_3 = \frac{\frac{f(x_3)-f(x_2)}{x_3-x_2} - \frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_3-x_0}$$

Pearson correlation .۲-۱-۱

ضریب همبستگی پیرسون، یک توصیف آماری است، به این معنی که ویژگی های یک مجموعه داده را خلاصه می کند. به طور خاص، قدرت و جهت رابطه خطی بین دو متغیر کمی را توصیف می کند. اگرچه تفسیر قدرت رابطه (همچنین به عنوان اندازه اثر نیز شناخته می شود) بین رشته ها متفاوت است، جدول زیر قوانین کلی را ارائه می دهد:

جدول ۱ - قوانین کلی برای ضرایب پیرسون

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

ضریب همبستگی پیرسون نیز یک استنباط آماری است، به این معنی که می توان از آن برای آزمون فرضیه های آماری استفاده کرد. به طور خاص، ما می توانیم آزمایش کنیم که آیا بین دو متغیر رابطه معناداری وجود دارد یا خیر.

راه دیگری برای در نظر گرفتن ضریب همبستگی پیرسون (r) به عنوان معیاری برای سنجش نزدیک بودن مشاهدات به خط بهترین تناسب نیز است. ضریب همبستگی پیرسون همچنین می گوید که آیا شیب خط بهترین تناسب منفی است یا مثبت. هنگامی که شیب منفی است، r منفی است. هنگامی که شیب مثبت است، r مثبت است.

- وقتی r = ۱ یا -۱ باشد، تمام نقاط دقیقاً روی خط بهترین تناسب قرار می گیرند.
- وقتی r بزرگتر از ۰.۵ یا کمتر از ۰.۵- باشد، نقاط به خط بهترین تناسب نزدیک هستند.
- وقتی r بین ۰ و ۰.۳ یا بین ۰ و ۰.۳- باشد، نقاط از خط بهترین تناسب فاصله دارند.
- وقتی r = 0 باشد، خط بهترین برازش برای توصیف رابطه بین متغیرها مفید نیست.

ضریب همبستگی پیرسون (r) یکی از چندین ضرایب همبستگی است که وقتی می‌خواهید یک همبستگی را اندازه‌گیری کنید، باید بین آن‌ها یکی را انتخاب کنید. ضریب همبستگی پیرسون زمانی انتخاب خوبی است که همه موارد زیر درست باشد:

- هر دو متغیر کمی هستند: اگر هر یک از متغیرها کیفی باشد، باید از روش دیگری استفاده کنید.
- متغیرها به طور معمول توزیع می شوند: می توانید یک هیستوگرام از هر متغیر ایجاد کنید تا بررسی کنید که آیا توزیع ها تقریباً نرمال هستند یا خیر. اگر متغیرها کمی غیر عادی باشند مشکلی نیست.
- داده ها هیچ نقطه پرت ندارند: نقاط پرت مشاهداتی هستند که از الگوهای مشابه بقیه داده ها پیروی نمی کنند. نمودار پراکندگی یکی از راه‌های بررسی نقاط پرت است - به دنبال نقاطی بگردید که از بقیه فاصله دارند.
- رابطه خطی است: "خطی" به این معنی است که رابطه بین دو متغیر را می توان به خوبی با یک خط مستقیم توصیف کرد. برای بررسی خطی بودن رابطه بین دو متغیر می توانید از نمودار پراکندگی استفاده کنید.

در زیر فرمولی برای محاسبه ضریب همبستگی پیرسون (r) آورده شده است:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

که در آن x و y نشان دهنده متغیرها و $\bar{x}$ و $\bar{y}$ نشان دهنده میانگین متغیرها هستند.

R^2 .۳-۱-۱

ضریب تعیین (Coefficient of determination) عددی بین ۰ و ۱ است که اندازه گیری می کند که چگونه یک مدل آماری یک نتیجه را پیش بینی می کند. ضریب تعیین اغلب به صورت R^2 نوشته می شود که به صورت "r مربع" تلفظ می شود. برای رگرسیون های خطی ساده، معمولاً از یک r کوچک به جای (r^2) استفاده می شود.

ضریب تعیین (R^2) اندازه گیری می کند که چگونه یک مدل آماری یک نتیجه را پیش بینی می کند. نتیجه با متغیر وابسته مدل نشان داده می شود. کمترین مقدار ممکن R^2، 0 و بالاترین مقدار ممکن ۱ است. به زبان ساده، هر چه یک مدل در پیش بینی بهتر عمل کند، R^2 آن به ۱ نزدیکتر خواهد بود.

- اگر $R^2 = 0$ باشد، مدل رگرسیون خطی به شما این امکان را نمی‌دهد که بهتر از تخمین ساده پیش‌بینی کنید.
- اگر R^2 بین ۰ و ۱ باشد، این مدل به شما امکان می دهد تا حدی پیش بینی کنید. تخمین های مدل کامل نیستند، اما بهتر از استفاده از میانگین هستند.
- اگر $R^2 = 1$ باشد، این مدل به شما امکان می دهد کاملاً پیش بینی کنید.

از نظر فنی تر، R^2 معیار خوبی برای تناسب است. این نسبت واریانس در متغیر وابسته است که توسط مدل توضیح داده می شود.

- مشاهدات به صورت نقطه نشان داده می شوند.
- پیش بینی های مدل (خط بهترین تناسب) به صورت یک خط سیاه نشان داده می شود.
- فاصله بین مشاهدات و مقادیر پیش بینی شده آنها (بقایای) به صورت خطوط بنفش نشان داده شده است.

ضریب تعیین نشان دهنده نسبت تمام تغییرات متغیر وابسته است که می تواند توسط متغیر مستقل از طریق رابطه رگرسیون توضیح داده شود. هر چه مقدار R2 به ۱ نزدیکتر شود، متغیر مستقل بهتر می تواند متغیر وابسته را توضیح دهد. که از فرمول زیر محاسبه می شود:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

در این سه معادله، n تعداد نمونه است، yi و y^i به ترتیب نشان دهنده مقدار واقعی و مقدار پیش بینی شده در زمان هستند. y¯i نشان دهنده میانگین همه مقادیر واقعی است.

۱-۲. دیتاست

دیتاست این مقاله حاوی اطلاعات هر ساعت از ۱۲سایت اندازه گیری آلاینده های هوا واقع در شهر Beijing چین میباشد.

توسط کتابخانه Pandas تمامی فایلهای excel فراخوانی شده است:

```
station_name = [i.split('_')[-2] for i in csv_files]
station_name
```

```
['Aotizhongxin',
 'Changping',
 'Dingling',
 'Dongsi',
 'Guanyuan',
 'Gucheng',
 'Huairou',
 'Nongzhanguan',
 'Shunyi',
 'Tiantan',
 'Wanliu',
 'Wanshouxigong']
```

```
# save all csv files in one dictionary
data = {}
for index, csv in enumerate(csv_files):
    data[station_name[index]] = pd.read_csv(csv,
                                    index_col=0,
                                    date_parser=lambda x: datetime.strptime(x, '%Y %m %d %H'),
                                    parse_dates=[['year', 'month', 'day', 'hour']])
```

```
data['Aotizhongxin']
```

year_month_day_hour	No	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	station
2013-03-01 00:00:00	1	4.0	4.0	4.0	7.0	300.0	77.0	-0.7	1023.0	-18.8	0.0	NNW	4.4	Aotizhongxin
2013-03-01 01:00:00	2	8.0	8.0	4.0	7.0	300.0	77.0	-1.1	1023.2	-18.2	0.0	N	4.7	Aotizhongxin
2013-03-01 02:00:00	3	7.0	7.0	5.0	10.0	300.0	73.0	-1.1	1023.5	-18.2	0.0	NNW	5.6	Aotizhongxin
2013-03-01 03:00:00	4	6.0	6.0	11.0	11.0	300.0	72.0	-1.4	1024.5	-19.4	0.0	NW	3.1	Aotizhongxin
2013-03-01 04:00:00	5	3.0	3.0	12.0	12.0	300.0	72.0	-2.0	1025.2	-19.5	0.0	N	2.0	Aotizhongxin

شکل ۱ - فراخوانی تمامی فایل های دیتاست

۱-۳. پیش پردازش

۱-۳-۱. Missing value

با استفاددده از تابع interpolate در دیتافریم پانداس و استفاده از متد linear داده های حذف شده را جایگزین میکنیم.

۱-۳-۲. Encoding Categorical Variable

طبق شکل زیر و مقادیر داده شده در مقاله، جهت باد را به درجه تبدیل میکنیم.

0°	— north wind	(N)
22.5°	— north-northeast wind	(NNE)
45°	— northeast wind	(NE)
67.5°	— east-northeast wind	(ENE)
90°	— east wind	(E)
112.5°	— east-southeast wind	(ESE)
135°	— southeast wind	(SE)
157.5°	— south-southeast wind	(SSE)
180°	— south wind	(S)
202.5°	— south-southwest wind	(SSW)
225°	— southwest wind	(SW)
247.5°	— west-southwest wind	(WSW)
270°	— west wind	(W)
292.5°	— west-northwest wind	(WNW)
315°	— northwest wind	(NW)
337.5°	— north-northwest wind	(NNW)
360°	— north wind	(N)

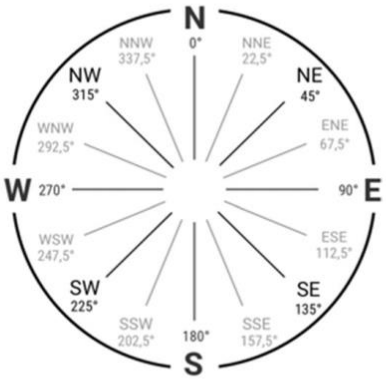


Fig. 7 Wind Direction and Degree Values

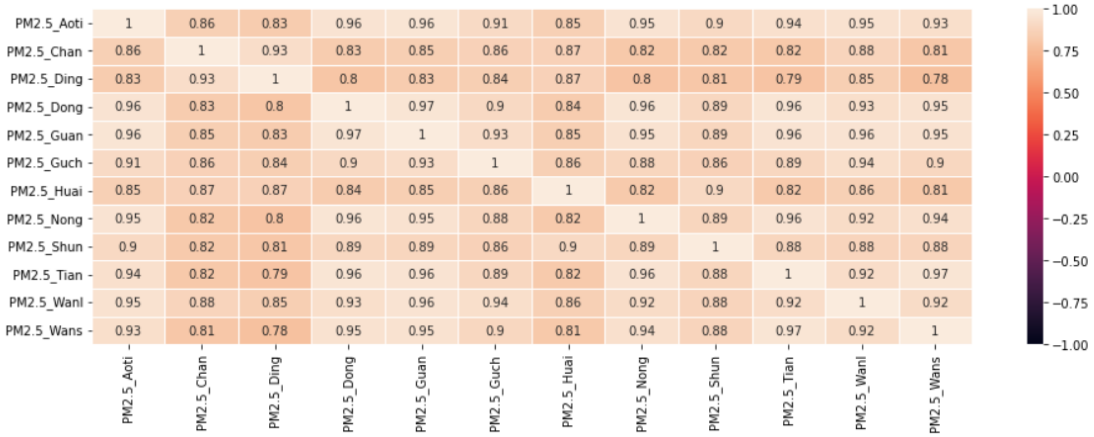
شکل ۲ - تبدیل جهت باد به درجه

۱-۳-۳. Normalization

با استفاده از تابع MinMaxScaler کتابخانه sklearn ، داده ها را به بازه [0, 1] فیت و سپس انتقال میدهیم.

۱-۳-۴. Pearson Correlation

با استفاده از دستور corr در دیتافریم پانداس و انتخاب متد pearson، pearson correlation را بدست آورده و بوسیله کتابخانه seaborn آن را بصورت heatmap نمایش میدهیم.که بصورت زیر درمی‌آید:



شکل ۳ - هیت مپ **pearson correlation** مربوط به **PM2.5** ایستگاه های مختلف

مشاهده میشود که ارتباط PM2.5 ایستگاه های مختلف زیاد است. بنابراین از آن ها به هنگام آموزش مدل استفاده میکنیم.

۱-۳-۵. Feature selection

دو دیتا فریم از PM2.5 ها و ویژگی ها مورد نظر ایستگاه Aotizhongxin بوسیله دستور concat به یک دیتافریم تبدیل و سپس درون فایل feature_result.csv ذخیره شده است.

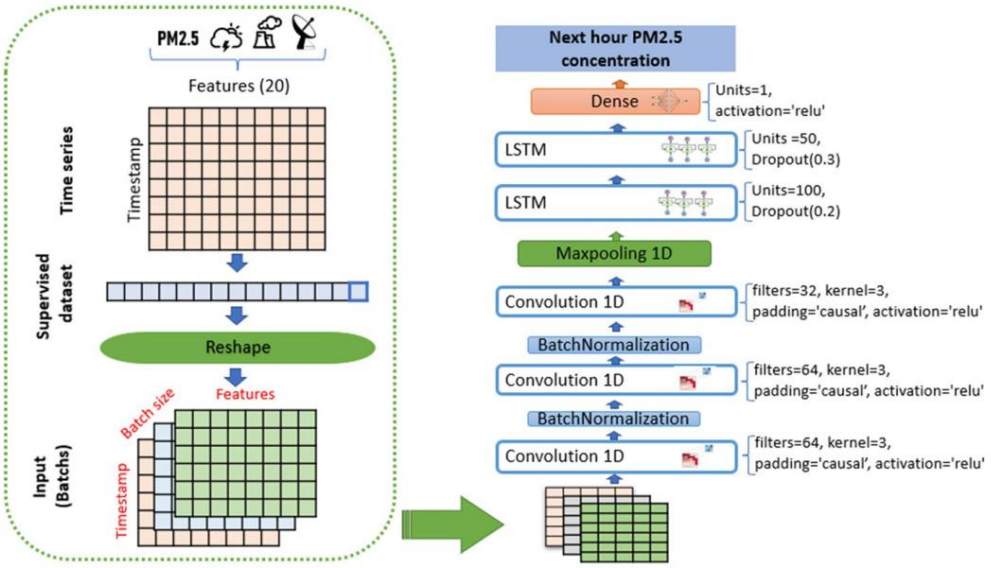
۱-۳-۶. Supervised dataset

داده ها را به این صورت supervised میکنیم که با توجه به lag که برابر ۱ یا ۷ روز باشد، به همان تعداد سمپل قبلی برای ورودی شبکه استفاده و خروجی آن را برابر سمپل بعدی قرار میدهیم.

به بیان دیگر بطور مثال برای lag = 7، چون تمام ساعات شامل ۱۶۸ = ۷*۲۴ میشود بنابراین از داده ۰ تا داده ۱۶۷ برای ورودی استفاده و داده ۱۶۸ را به عنوان لیبل و خروجی در نظر میگیریم. به همین ترتیب برای هر سمپل خروجی یعنی در هر ساعت با توجه به lag به همان تعداد سمپل قبلی را بعنوان ورودی انتخاب میکنیم که کد آن زده شده است.

۱-۴. آموزش شبکه

پس از پیش پردازش داده ها و supervised کردن آن ها، مدل CNN-LSTM بکار رفته در مقاله را باتوجه به عکس زیر ایجاد نمودیم:



شکل ۴ - معماری **CNN-LSTM** بکار رفته در مقاله

سپس مدل را با داده های آموزشی ایجاد شده، با تابع های loss، MAE و MSE بصورت جداگانه برای هر lag = 1, 7 به تعداد epoch = 25 برای batch = 32 آموزش داده شده است.

در این آموزش مدل، از الگوریتم Adam برای بهینه سازی استفاده شده است، که نتایج آن که شامل ۳ بخش است در زیر مشاهده میشود:

۱- در این مدل از ۳ متریک مختلف، mean square error, mean absolute error, coefficient correlation استفاده شده است که مقادیر آن ها بر روی داده های تست و آموزش در عکس اول هر بخش میباشد:

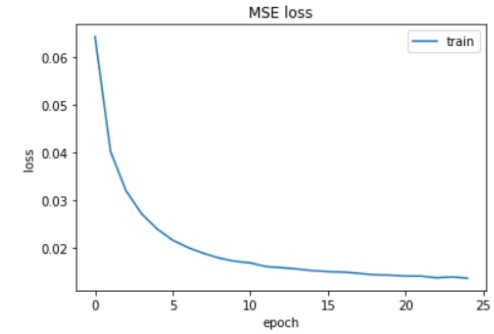
۲- در شکل دوم نمودار loss بر روی داده های آموزش نشان داده شده است:

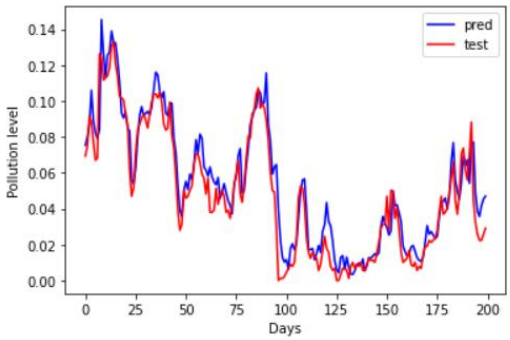
۳- در شکل سوم هر قسمت نیز نتایج تست و نتایج پیش بینی شده مقدار آلودگی هوا مقایسه شده است:

۱- برای lag = 7:

- MSE:

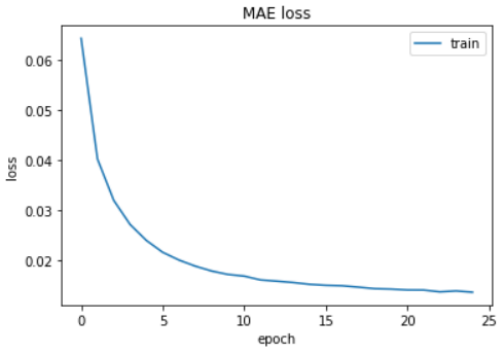
```
Epoch 25/25
876/876 [==============================] - 13s 14ms/step - loss: 3.7946e-04 - mean_absolute_error: 0.0127 - mean_squared_error:
3.7946e-04 - r2_score: 0.9484 - val_loss: 3.9836e-04 - val_mean_absolute_error: 0.0126 - val_mean_squared_error: 3.9836e-04 - v
al_r2_score: 0.5613
```

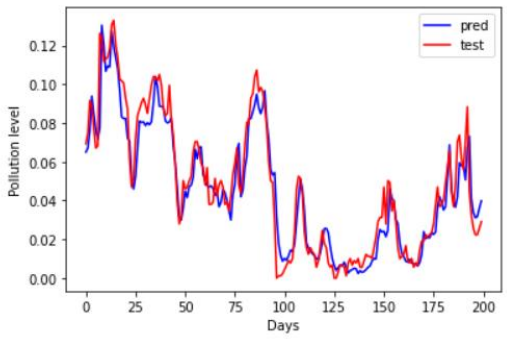




- MAE:

```
Epoch 25/25
876/876 [==============================] - 13s 14ms/step - loss: 0.0136 - mean_absolute_error: 0.0136 - mean_squared_error: 5.1
438e-04 - r2_score: 0.9292 - val_loss: 0.0120 - val_mean_absolute_error: 0.0120 - val_mean_squared_error: 3.9998e-04 - val_r2_s
core: 0.6585
```

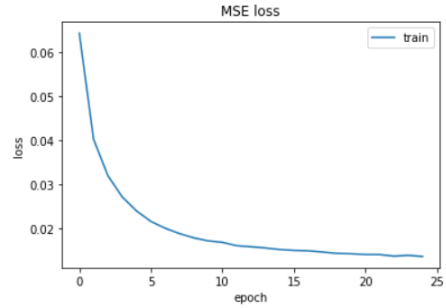


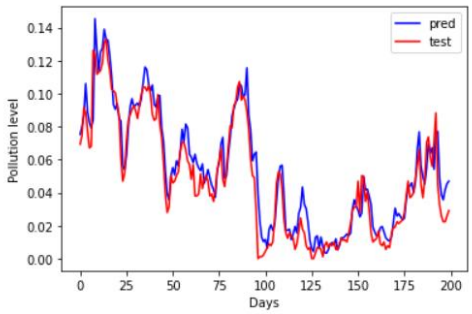


۲- برای lag = 1:

– MSE:

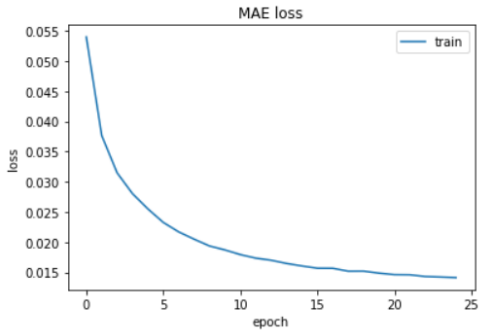
```
Epoch 25/25
876/876 [==============================] - 13s 14ms/step - loss: 3.7946e-04 - mean_absolute_error: 0.0127 - mean_squared_error:
3.7946e-04 - r2_score: 0.9484 - val_loss: 3.9836e-04 - val_mean_absolute_error: 0.0126 - val_mean_squared_error: 3.9836e-04 - v
al_r2_score: 0.5613
```

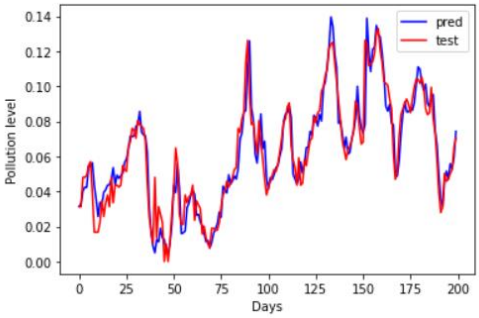




– MAE:

```
Epoch 25/25
876/876 [==============================] - 10s 11ms/step - loss: 0.0141 - mean_absolute_error: 0.0141 - mean_squared_error: 5.6
125e-04 - r2_score: 0.9271 - val_loss: 0.0135 - val_mean_absolute_error: 0.0135 - val_mean_squared_error: 4.4017e-04 - val_r2_s
core: 0.4586
```





بررسی مقایسه ای MAE و RMSE نشان می دهد که نه تنها کمترین میانگین مطلق خطا، بلکه کمترین ریشه میانگین مربع خطا نیز در مدل پیشنهادی رخ می دهد.