Machine Learning Techniques
Assignment: Air Quality Forecasting
Student: Nhial Majok Riak
Date: 9/21/2025

1. Introduction

Heavy air pollution is a long-standing problem in China's capital due to industrialization, urbanization and widespread use of vehicles. PM2. 5 is a top component of air pollution since it is so small it can go deep within the body and get into the blood system leading to both respiratory as well as cardiovascular disease. Accurate forecasting of PM2. 5 levels, the researchers say its important for policymakers, environmental agencies and people to take action before it's too late.

Classical statistical models have been widely applied to the time series forecast, such as Autoregressive Integrated Moving Average (ARIMA) model. Nonetheless, these models often struggle with modeling the intricate timing structures and nonlinear interactions of pollution data. During the last few years, machine learning models, especially deep learning models such as RNNs, LSTM and GRU, have proved to be very promising in solving these problems. These models are tailored to sequential data and can solve the problem of learning long-term dependencies, thus making them well suited for air quality prediction.

Objectives:

- Explore and preprocess sequential data.
- Implement LSTM/GRU/CNN-LSTM models.
- Vary the hyperparameters to minimize error.
- Performance may be assessed by Root Mean Squared Error (RMSE)
- Submit predictions to Kaggle.

Target: RMSE<4000 on the Kaggle leaderboard.

2. Data Exploration and Preprocessing

Dataset Summary

- Train: 30,676 samples, 11 features + target pm2. 5.
- Test: 13,148 samples, 10 features (no target).
- Attributes: Dew point (DEWP), Temperature (TEMP), Pressure (PRES), Wind speed (Iws), Categorical wind direction (encoded: cbwd_NW, cbwd_SE, cbwd_cv), etc.

Missing Values

- pm2.5: 6.26% missing.
- Tested imputation methods:
  - Mean → RMSE = 88.82
  - Forward/backward fill → RMSE = 22.52
  - Time interpolation (best) → RMSE = 15.35
  - kNN → RMSE = 22.51

Final choice: Time interpolation.

Feature Engineering

- Dropped No column.
- Standardized continuous variables (mean=0, std=1).
- Encoded wind direction with one-hot.
- Built sliding windows of 24–72 hours.

Resulting shapes:
Train: (27,586, 24, 9)
Validation: (3,066, 24, 9)

3. Model Design

I developed and evaluated four architectures:

1. Baseline LSTM
   1 layer, 64 units, dropout=0.15 Adam
2. GRU
   64–128 units, less parameters, faster convergence.
3. Bidirectional LSTM (BiLSTM)
   Captures forward and backward dependencies.
4. CNN-LSTM
   Applies Conv1D+LSTM for local time pattern learning.

Loss: Mean Squared Error (MSE)
Metric: RMSE

4.Experiment Table

Saved experiment results to: /content/experiment_grid_results.csv

	exp	tag	seq_len	arch	lr	batch_size	val_rmse	val_mae	val_r2	time_s
0	15	GRU_L48_lr0.0008	48	GRU	0.0008	128	70.136931	45.940719	0.562093	62.775022
1	6	GRU_L48_lr0.002	48	GRU	0.0020	128	71.076437	46.641562	0.550283	40.388762
2	9	LSTM_L72_lr0.002	72	LSTM	0.0020	128	71.524062	47.834712	0.543699	51.793367
3	10	GRU_L72_lr0.002	72	GRU	0.0020	128	72.256855	46.608847	0.534301	63.351544
4	14	GRU_L48_lr0.0012	48	GRU	0.0012	128	72.887655	47.095157	0.527071	35.506926
5	13	GRU_L48_lr0.003	48	GRU	0.0030	128	74.547606	47.511381	0.505285	45.716253
6	11	BiLSTM_L72_lr0.0015	72	BiLSTM	0.0015	128	75.132938	47.743522	0.496490	68.064272
7	2	GRU_L24_lr0.002	24	GRU	0.0020	128	76.247139	48.512899	0.482903	40.508276
8	5	LSTM_L48_lr0.002	48	LSTM	0.0020	128	77.754253	49.282836	0.461809	31.652916
9	7	BiLSTM_L48_lr0.0015	48	BiLSTM	0.0015	128	77.803852	50.761714	0.461122	51.235578
10	4	CNN-LSTM_L24_lr0.0015	24	CNN-LSTM	0.0015	128	77.860083	51.269188	0.460794	26.174927
11	3	BiLSTM_L24_lr0.0015	24	BiLSTM	0.0015	128	79.533975	51.036009	0.437361	36.255307

Best experiment summary:

{'rmse': 70.13693077463957, 'mae': 45.9407186572525, 'r2': 0.5620934366689957, 'seq_len': 48, 'arch': 'GRU', 'lr': 0.0008, 'model': <Sequential name=sequential, built=True>

5. Results & Discussion

RMSE Definition

$RMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ RMSE= n 1i=1 ∑ n (y i− y ^ i) 2

Performance

- Validation RMSE: 51.8872901160902 (best GRU model).
- Kaggle private score 4205.6987 (over target 3000).

Error Analysis

- Low validation RMSE with higher Kaggle test RMSE → probably distribution shift.
- Pollution peaks (spikes) were also underpredicted.
- (8) BiLSTM & DaCNN-LSTM Not surprisingly, both of them perform slightly better than their LSTM-C counterparts.

Challenges

- Training time for deep RNNs.
- Missing data imputation.
- Potential to vanish gradients (mitigated by GRU + dropout).

6. Conclusion

In this study, deep learning was used to predict PM2. 5 concentrations.
Key takeaways:

- For the 72 hour sequences, GRU with learning rate 0.0005 exceeded other models.
- Interpolation through time fared better than imputation vs mean/ffill.
- Final Validation RMSE - 51.8872901160902, Kaggle 4205.6987(Should have done better).

Future directions:

- Apply attention-based LSTMs/Transformers. Use models based on the ensemble of GRU and CNN-LSTM.
- Add external factors (public holidays, traffic, weather forecasts).
-

7. GitHub Repository

Code, notebook, logs can be found here:
https://github.com/nmaketh/air-quality-forecasting
8. References (IEEE style)

[1] Shaojun Zhang, et al_"Progress of Air Pollution Control in China and Its Challenges and Opportunities in the Ecological Civilization Era." Engineering, vol. 6, no. 12, 2020, pp. 1423-1431. . [Online serial]. Available: https://www.sciencedirect.com/science/article/pii/S2095809920301430. [Accessed Dec. 2, 2020]].

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[3 L. Li, Q. Wang, T. Zhang, J. Huang, and J. Fan, "Air quality forecasting using hybrid deep learning method based on CEEMDAN and BiGRU," Environmental Technology & Innovation, vol. 20, p. 101092, 2020, doi: 10.1016/j.eti.2020.101092.