Air Quality Multistep Forecasting - Report

Overview:

The goal of this project is to develop a robust multistep forecasting model for air quality prediction. The project encompasses data preprocessing, model development using Bidirectional Gated Recurrent Units (Bi-GRU), training, testing, and evaluation. The dataset consists of historical air quality data for multiple cities.

Model Architecture:

The chosen architecture revolves around Bidirectional Gated Recurrent Units (Bi-GRU), a type of recurrent neural network designed to capture temporal dependencies in sequential data. The model architecture is as follows:

```
_________________________________________________________________
Layer (type)                Output Shape              Param #
=================================================================
bidirectional (Bidirectional (None, 120, 256)         124416

_________________________________________________________________
dropout (Dropout)           (None, 120, 256)          0

_________________________________________________________________
bidirectional_1 (Bidirection (None, 128)              123648

_________________________________________________________________
dropout_1 (Dropout)         (None, 128)               0

_________________________________________________________________
dense (Dense)               (None, 32)                4128

_________________________________________________________________
dropout_2 (Dropout)         (None, 32)                0

_________________________________________________________________
dense_1 (Dense)             (None, 16)                528

_________________________________________________________________
dense_2 (Dense)             (None, 1)                 17
=================================================================
Total params: 252,737
Trainable params: 252,737
Non-trainable params: 0

_________________________________________________________________
```

The model includes Bidirectional GRU layers with dropout for regularization and Dense layers for output. This architecture aims to capture intricate patterns in air quality time series data.

Training Process:

The training process involves the following steps:

Data Normalization: Features and target data are normalized using MinMaxScaler.

Sequence Generation: Sequences for multistep forecasting are created using a defined history size and target size.

Model Compilation: The model is compiled with the Adam optimizer and Mean Squared Error (MSE) loss function.

Model Training: The model is trained on the training data with validation on a separate validation set.

Checkpoints and Early Stopping: ModelCheckpoint and EarlyStopping callbacks are implemented to save the best model and prevent overfitting.

Testing Process:

The testing process involves the following steps:

Data Loading: Test data is loaded for each city.

Normalization: Test features are normalized using the same scalers used during training.

Sequence Creation: Sequences for testing are generated.

Model Prediction: The trained model is used to make predictions.

Metric Calculation: Metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) are calculated for evaluation.

Evaluation Results:

Evaluation results are saved in the 'results' folder, including sample graphs and metrics for each city and step size. The metrics provide insights into the model's performance in terms of accuracy and generalization.

Model Summary:

The model is summarized with detailed layer-wise information, providing insights into the architecture's complexity and parameters. The use of Bidirectional GRU layers enhances the model's ability to capture temporal dependencies from both past and future timestamps, making it suitable for multistep forecasting.

This project showcases an end-to-end workflow for air quality multistep forecasting, combining robust architecture, effective training strategies, and thorough evaluation processes. The use of Bidirectional GRU layers contributes to the model's success in capturing complex patterns within the temporal nature of air quality data.