Deep Learning - Assignment 2

Group 2

Instructor:
Siamak Mehrkanoon

Group Members:
Sheng Kuang (i6237193)
Yimin Yang (i6246099)

---

1. Brief answers for the questions

Q1: Which model do we use?

A1: LSTM and **CNN** (Section 2 and Section 3).

Q2: What is the mean absolute error of predictions of each model?

A2: LSTM-5: 2.903, **LSTM-6: 1.466 (Optimal)**, CNN: 2.100. (Section3, Figure 2).

Q3: Which features among the 5 measured features are more relevant for outputs of the model?

A3: 'Temperature' (Section 4.2).

Q4: How many previous steps do we need to feed into the model?

A3: LSTM: 100 (Occlusion analysis: the recent **45 hours** are the most important, especially the recent **10 hours**.); **CNN: 80** (Section 3.3).

2. Motivation of Model Selection

2.1 LSTM Weather forecasting requires the model to handle sequential data. Usually, the dataset for weather forecasting consists of records with several features with a sliding window. To keep temporal information, LSTM is selected as our first model to predict the temperature.

2.2 CNN Even though LSTM is very good at predicting series problems, CNN is also another choice. While treating two LSTM models as our mainstream solution, we also use CNN to train the model and make some simple comparisons of the results as the extension to the assignment.

3. Method

3.1 Pre-processing

3.1.1 Hypothesis Assuming the weather of different cities is independent, the correlations of weather between different cities are not taken into consideration in our assignment.

*** we also try to combine the correlations into the model but the results of occlusion analysis are kind of unexplainable***

3.1.2 Normalization The normalization approach we selected is ***MinMaxScaler***. The range of scaled data is from 0 to 1. The first two dimensions of the raw data (records*cities*features) are reshaped to ensure all the data of each feature are scaled together.

3.1.3 Training & Validation Dataset Excluding the last 168 records for testing data, the remaining 90% of the data are used for the training set and the remaining 10% are used for the validation set.

3.2 Architectures

Two LSTM and one CNN models are investigated in our approach. The main difference between the two LSTMs is the feature numbers.

3.2.1 LSTM with 5 Features (LSTM-5) The model consists of an LSTM layer and a linear layer on the top of it. The LSTM layer contains 100 units, 5 hidden states for each unit, no stack layers, and no bidirectional propagation. The input size is [batch, lags, features] while the output size equals [batch, 1].

3.2.2 LSTM with 6 Features (LSTM-6) In 3.2.1, the temperature predictions of different cities are based on the same model. However, the weather in different locations might be totally different due to geographical impact. When this information is missing, we introduce a new feature, called 'city index', to make it possible for the model to predict temperatures in different cities based on different feature weights. The 'city index' feature is the index of the second dimension in raw data. Thus, the input embedding size is set to 6. Other model settings are keeping the same.

3.2.3 CNN The CNN model we use is basically the same as the CNN model we built for assignment 1. The model consists of a convolutional layer, a max pooling layer, and a Flatten layer followed by two Dense layers. The main difference is we used Conv1D in the convolution layer last time, and this time we use Conv2D. This also happens for the pooling layer.

3.3 Experiment Setting Up

The best hyperparameter combination of each model is based on the fine-tuning experiments. For all the experiments, the learning rate (LR) is set to 0.001 to minimize mean square error(MSE). In addition to LR, the following hyperparameters and settings are investigated:

- Optimizer: SGD (Momentum), RMSProp, Adagrad, Adam
- Lags: ranges from 10 to 100
- Batch Size: ranges from 32 to 1000
- Hidden Size (LSTM): ranges from 4 to 14
- Num Layers (LSTM): ranges from 1 to 3

The convergence speed of three models is shown in Appendix A and B.

3.4 Occlusion Analysis

Occlusion analysis is a method to compare the importance of features after training. In this assignment, we investigate spatial occlusion analysis to explore the importance of features to cities, and temporal occlusion analysis to explore the time lags. The rationale is to compare the MSE percentage change between reference MSE and a new MSE, obtained by prediction of masked features[1]. The change corresponds to the importance of features. When masking a feature of the input data to the model, the true values are replaced by values from a random uniform distribution between 0 and 1 (corresponds to the MinMaxScale in 3.1).

4. Results

4.1 Model Comparison

The MAE of temperature prediction of aforementioned models are presented in Table 1. The average MAE of 4 cities prediction by LSTM-6 is 1.518C, which is the optimal performance. The MAE of 'City 3' temperature prediction by LSTM-6 is 1.466C while the best lags is 100. The recursive prediction of 'City 3' is shown in Figure 1.

Table 1. The MAE comparison of three models

Model	City Index	MAE (C)	AVG MAE (C)
LSTM-5	0	2.143	2.766
	1	2.327	
	2	3.692	
	3	2.902	
LSTM-6	0	1.080	**1.518**
	1	1.641	
	2	1.885	
	3	**1.466**	
CNN	0	2.575	2.308
	1	2.215	
	2	2.340	
	3	2.100	

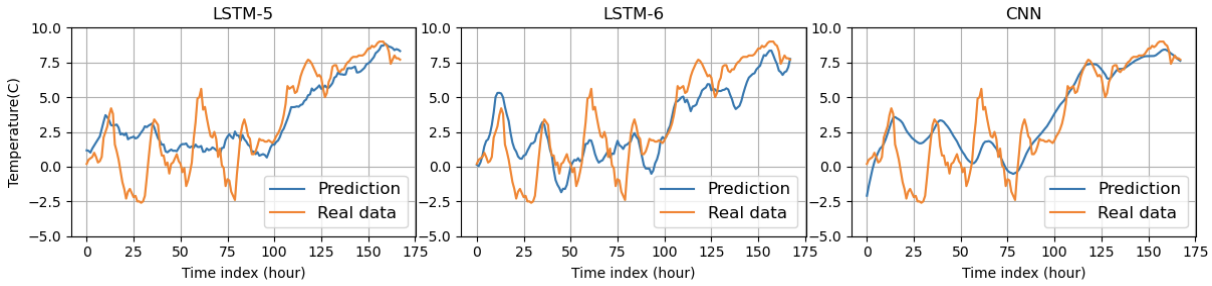


Figure 1. Temperature prediction of 'City 3' by LSTM-5, LSTM-6 and CNN.

4.2 Occlusion Analysis

The target model of the occlusion analysis is LSTM-6. The percentage change of MSE is the average percentage change for 10 times because feature values are replaced by a random uniform distribution.

4.2.1 Features The obtained MSE percentage change for all features is listed in Table 2. Figure 2(a) is the visualization of obtained results. From Table 2 and Figure 2, 'Temperature' itself is the most important feature of each city. For 'City 0', 'City 1' and 'City 2', 'Dew Point' is the second most important feature while 'Wind Speed' is the second for 'City 3'. Meanwhile, 'Wind direction' has an extremely low impact on each city. The feature importance of temperature forecasting of the 4 cities is relatively similar.

Table 2. The occlusion analysis of LSTM-6 illustrates the importance of each feature to each city

Feature Name	Δ MSE (%)			
	City 0	City 1	City 2	City 3
Wind Speed	19.27	1.56	-5.67	201.30
Wind Direction	-16.39	-8.73	-12.51	-5.34
Temperature	**752.09**	**777.66**	**682.40**	**1087.30**
Dew Point	256.46	34.78	30.65	190.98
Air Pressure	17.35	7.32	1.38	110.85

4.2.2 Lags We add a sliding window to the mask to explore the lag influence. The mask sizes are '5*1', '10*1', '5*6', and '10*6' (i.e., 5/10 is the lags). The size of '5*1' and '10* 1' means only one feature is masked with 5/10 lags in every experiment, while the size of 5*6 and 10*6 means all features are masked with 5/10 lags. Figure 2.(b) shows the impact of lags in different features. The recent lags (right side) have more significant impact on predictions when compared to old lags (left side). Observing the bright area of figures, the recent 45-hour temperature plays a crucial role in the model, especially the recent 10 hours. It is also worth mentioning that recent 30-40 hour lags are more important than recent 20-30 hour lags. Figure 2.(c) indicates the performance when features are simultaneously masked in a certain time window. The temperature prediction of 'City 3' is particularly sensitive to the last 5 hours with a resolution of 5 (i.e., the mask length equals 5). However, at a resolution of 10, 'City 0' is more sensitive to the last 10 hour lags than any other city.

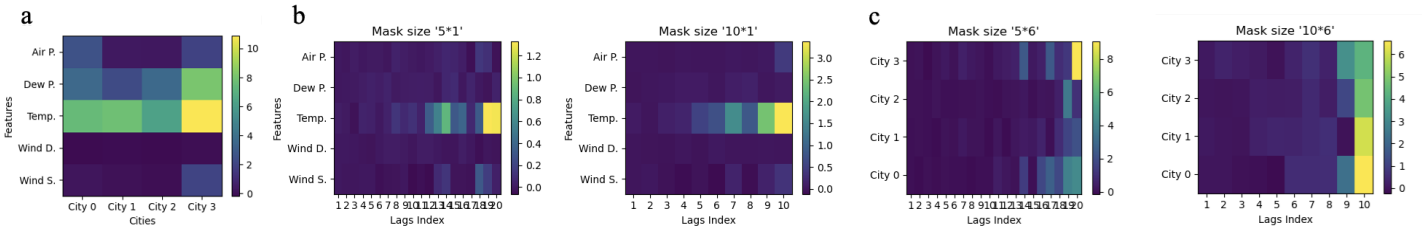


Figure 2. The spatial and temporal occlusion analysis. a) Spatial analysis shows the importance of features to cities. b) Temporal analysis indicates the impact of lags on temperature prediction when features are masked respectively. c) Temporal analysis indicates the lag influence of each city when features are simultaneously masked by a 5*6 or 10* 6 size mask.

5. Conclusion

In this assignment, three kinds of models are investigated to predict the temperature.The LSTM model with an additional city feature has the best performance. Occlusion analysis is used to find the most important features to the model. When predicting temperature, the recent 45 hour ‘Temperature’ data are the most important for the model, especially the recent 10 hours.

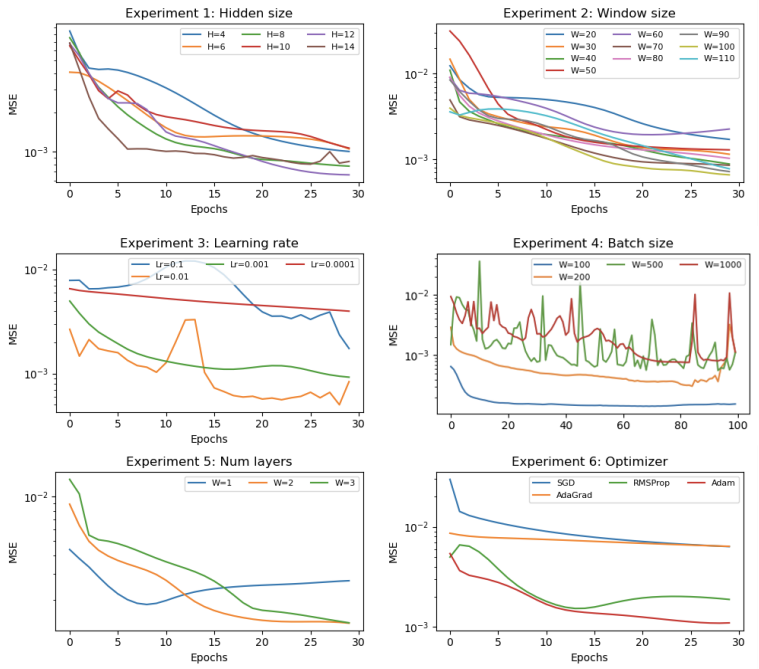
6. Reference

[1] Abdellaoui, Ismail Alaoui, and Siamak Mehrkanoon. "Deep multi-stations weather forecasting: explainable recurrent convolutional neural networks." *arXiv preprint arXiv:2009.11239* (2020).

[2] Tekin, S. F. et al. “Spatio-temporal Weather Forecasting and Attention Mechanism on Convolutional LSTMs.” ArXiv abs/2102.00696 (2021): n. pag.

Appendix

A. LSTM convergence speed with different hyperparameters



B. CNN convergence speed with different hyperparameters