Assessing Methods for Handling Missing Data Using an LSTM Deep Learning Model in Traffic Forecasting

Author: **Wouter Büthker**
linkedin.com/in/wbuthker w.w.buthker@student.tudelft.nl
Responsible professor: Elena Congeduti
Examiner: George Iosifidis

1. Background

- Accurate traffic prediction can improve transportation efficiency, reduce traffic congestion, and create safer roads [1].
- Unfortunately, real-world data often contain inaccuracies due to reasons like sensor failure or communication errors. This is not different in the traffic domain.
- Research into traffic forecasting has shown the use of Long Short-Term Memory (LSTM) [2] outperforms traditional time-series models and shallow neural networks [3]

2. Research Questions

1. How much does missing data affect the accuracy of an LSTM traffic prediction model?
2. What is an effective strategy to improve model performance using erroneous data?

3. Data analysis

- There's only a small amount (0.34%) of data missing from the source data set.
- Missing values appear in two patterns:
  1. Sequences at random intervals.
  2. Single missing values at random or regular intervals.
- This research focuses on the first pattern.
- The data set used in this paper contains is collected by the municipality of The Hague. Induction loops, often in front of traffic lights, were used to count the number of vehicles passing per 15 minutes, the traffic flow.

4. Methodology

Data points were manually removed from complete datasets, mimicking the patterns of missing data found in the original data.
Given a set of historical traffic flow observations $S = (x_1, x_2, \dots, x_n)$ where x_i denotes the traffic flow at time step i, we create a subsequence $X = (x_s, x_{s+1}, \dots, x_t)$, where $X \subset S$. The aim is to predict the traffic flow y_{t+1} at a future time step $t + 1$.
Three different methods to handle missing data are explored

1. Drop null values: $S'_p = (x_1, x_2, \dots, x_{t-p})$ where p is the amount of $null$ values.
2. Set null values to zero $S'_p = (x_1, x_2, \dots, x_t)$ where p values of x have been replaced with 0.
3. Linear interpolation $S'_p = (x_1, x_2, \dots, x_t)$ where p values of x have been interpolated

Regardless of the strategy, the LSTM is trained on subsequences $(X_1, \dots, X_n)$ based on the artificial data set S'. Finally, the model is evaluated against a complete baseline data set S. Multiple different values for p are used to determine how the amount of missing data affects each strategy.

6. Conclusion

- Based on the results it can be concluded that surprisingly, even for high amounts of missing data the model is still able to make fairly good predictions. Both interpolation and setting values to zero lead to a small predictable increase in RMSE.
- For less than 40% of the data missing, the choice of strategy to handle missing data does not have much impact. In the case where there is more data missing, dropping values should be avoided, since it proved to be less reliable than the other strategies.

7. Further work and Limitations

- Further work should investigate model performance when the model is trained with a perfect data set but fed missing data.
- Besides missing values, source data can also contain unreasonably values. These kind of errors are not explored further in this research. Further work could investigate how these kinds of errors affect a model.

5. Results

The results demonstrate the surprising resilience of LSTM models to missing data in Figure 1. For up to 40% missing data, there is little impact on the prediction accuracy, regardless of the strategy used. For higher proportions of missing data, the strategy of dropping null values significantly degrades performance, while zero-filling and interpolation maintain reasonable predictive accuracy, with the RMSE only increasing slightly.

Figure 2, 3 and 4 show a random sample of 5 days of the traffic flow at a random location. Figure 2 and 3 use a model trained using method 2 when removing 95% of the data. In Figure 2 the training set is visible with the model's predictions. Figure 3 and 4 show predictions on the same test set. The model from Figure 3 is trained on a data set with only 5% data intact while Figure 4 has been trained on a perfect data set. This comparison shows us that even with little data the model is still able to predict the traffic flow well.

Figure 1: Performance of LSTM using different amounts of missing data, for the three strategies and a baseline.

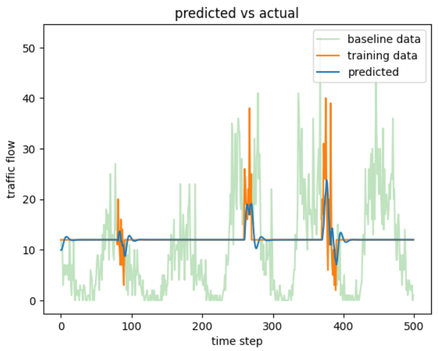

Figure 2: Model predictions using a model trained on 5% of the original dataset. A sample of 5 days from the training set is shown.

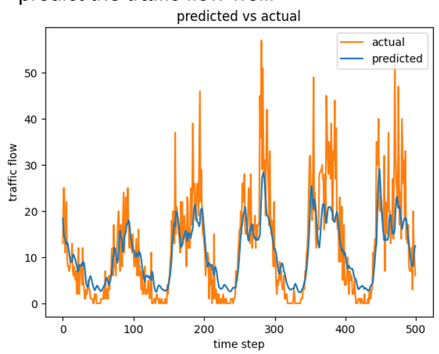

Figure 3: Model predictions using a model trained on 5% of the original dataset. A sample of 5 days from the test set is shown.

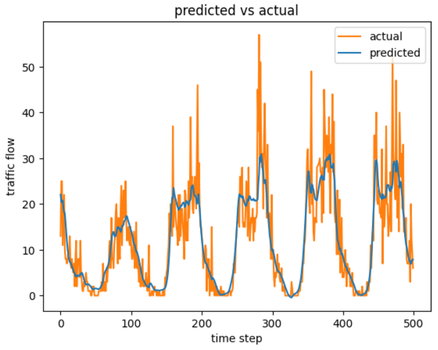

Figure 4: Model predictions using a model trained on the unmodified dataset. A sample of 5 days from the test set is shown.

8. References

[1] S. Foo and B. Abdulhai. Evaluating the impacts of changeable message signs on traffic diversion. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 891–896, 2006
[2] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.
[3] Yisheng Lv, et al. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.