Using LSTMs To Predict Carpark Occupancy Rates in Singapore

Chua Zong Wei (A0183028H),[1] Marcus Duigan Xing Yu (A0199347L), [1]
Munnamgi Harsha Vardhan Reddy (A0200162A), [1] Teh Nian Fei (A0171124U), [1]
Teo Hoe Keat (A0200203L), [1] Zou Runzhong (A0206055N)[1]

National University of Singapore (Group 11)[1]
e0309823@u.nus.edu[1]

Abstract

In Singapore where land space use is a challenge, space dedicated to carparks is projected to decrease as vehicular traffic increases. To better optimize the use of current and future carparks, an LSTM model was trained to predict carpark occupancies of seven carparks around Serangoon given past occupancy rates, weather conditions, and other time-derived features. The LSTM model trained achieved a low validation MSE of 0.001 across all seven carparks, which is sufficiently accurate to predict future carpark occupancies to the nearest integer. It was also shown to outperform alternative models such as gradient-boosted trees. However, several limitations of the model, as well ethical concerns with using it for applications such as dynamic pricing were also identified. These issues need to be sufficiently addressed before the model is used in real-world settings.

Introduction

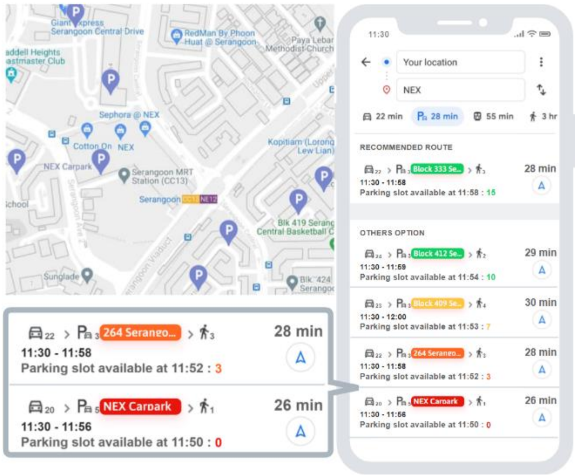
With decreasing amounts of available land space in the small island nation of Singapore, it is foreseeable that less land space can be dedicated to carparks in the near future. This may pose a challenge as carpark spaces become more and more difficult to find as the number of cars slowly increase in Singapore. Moreover, there is also an uneven distribution of carpark availabilities in different areas, or even across different carparks in the same area. Carpark space availability may also vary greatly depending on various factors such as the time of day and weather. Thus, there is a need to predict the availability of carpark spots based on data that potentially affects availability. A machine learning model is suitable for this as it is difficult to pinpoint the factors that affect carpark occupancy, given that these factors may themselves be correlated with other factors.

Potential Applications

There are several important applications for observing trends in the carpark occupancy. The data generated through can be useful for individual drivers to optimize the time required to find parking. It has been found that an average Singaporean spends 19 minutes to find parking (Kee, 2020), especially during peak hours. Typically, drivers are not aware of available alternative carparks and have no choice but to wait for a parking spot in the carpark that they are currently in to vacate. Although there already exist databases that provide current carpark occupancy data, the occupancy of a carpark may change by the time the driver arrives at their destination, which makes the current carpark occupancy inadequate for planning trips ahead of time. The accurate projection of future carpark occupancy rates can thus inform drivers ahead of time regarding which carparks in an area will have free spots upon their arrival.

Using future carpark occupancies, suggestions on which carparks are likely to have more spaces based on estimated time of arrival, weather at destination, and other factors can be generated. This reduces the amount of time spent looking for parking. With less time spent on this, not only can consumers save on fuel expenditure, but Singapore’s carbon emissions can also be reduced. Since vehicle emissions are the third largest contributor of Singapore’s carbon emissions (NCCS, 2017), Singapore’s carbon footprint can be significantly reduced by reducing vehicular emissions.

Using data regarding predicted future peak periods, dynamic pricing for carparks can also be introduced. Like Electronic Road Pricing (ERP), dynamic parking rates can help further manage congestion for carpark use by encouraging drivers to arrive during non-peak timings to have cheaper parking. Furthermore, with the rise of autonomous vehicles, such models could be integrated directly into the software such that the decisions are made automatically depending on the driver’s preferences and willingness to pay. Urban planning authorities would also be able to better forecast the level of congestion with the integration of such software in all vehicles.


Giant Express Serangoon Central Drive
RedMan By Phoon Huat @ Serangoon
Paya Lebar Methodist Church
addell Heights astmaster Club
Sephora @ NEX
Cotton On
NEX
NEX Carpark
Serangoon MRT Station (CC13)
Kopitiam (Lorong Lew Lian)
Serangoon
Blk 419 Serangoon Central Basketball C
Sunglade
Blk 424 Serangoon

22 › 3 264 Serango... › 3 28 min
11:30 - 11:58
Parking slot available at 11:52 : 3

20 › 5 NEX Carpark › 1 26 min
11:30 - 11:56
Parking slot available at 11:50 : 0

11:30
Your location
NEX
22 min | 28 min | 55 min | 3 hr

RECOMMENDED ROUTE

22 › 3 Block 333 Se... › 3 28 min
11:30 - 11:58
Parking slot available at 11:58 : 15

OTHERS OPTION

24 › 3 Block 412 Se... › 2 29 min
11:30 - 11:59
Parking slot available at 11:54 : 10

23 › 3 Block 409 Se... › 4 30 min
11:30 - 12:00
Parking slot available at 11:53 : 7

22 › 3 264 Serango... › 3 28 min
11:30 - 11:58
Parking slot available at 11:52 : 3

20 › 5 NEX Carpark › 1 26 min
11:30 - 11:56
Parking slot available at 11:50 : 0

Figure 1: Parking lots available near the destination (left) and example of carpark routing information incorporated into a popular navigation application, Google Maps (right).

The model's results can be in incorporated into navigation applications to help drivers find available parking lots near their destination. With projected carpark vacancy data, the navigation application can search carparks near the driver's destination and use this information to estimate the time that would be required to search for an empty lot at each carpark. Like modern navigation applications, it can present several choices balancing between travel time to the carpark, time required to find an empty lot, and walking time to the final destination, among other factors (Fig. 1). The GPS navigation interface can also be enhanced with not just real-time carpark vacancy information, but projected vacancy information at the time of arrival at said carpark (Fig. 2).

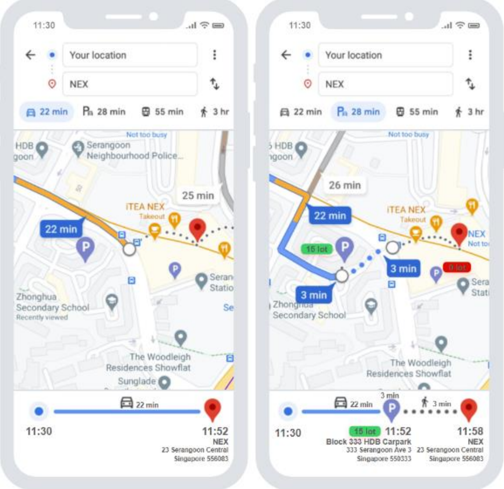


Figure 2: Google Maps without carpark routing information (left) and enhanced with carpark routing information (right).

The model can be hosted on a centralized server, as carpark occupancy rate predictions will be the same for all users. However, real time streaming data regarding each user's intended destination carpark after viewing the model's prediction should ideally be fed back to the model, since the model's prediction at a current time also has an impact on the occupancy rate of the carpark in the future.

Long-Short Term Memory (LSTM) Networks

Addressing the problem statement as defined earlier requires a machine learning framework capable of exploiting the non-linear structure of the processed dataset. In particular, the model has to be capable of capturing sequential dependencies, in which the current output is dependent on the previous input.

One of the models initially identified was a Recurrent Neural Network (RNN), due to its temporal dynamic behaviour, which was suitable to model sequential data (Shekhar, 2019).

The most basic RNNs are variants of deep neural networks, whose constituent feedforward neural networks are capable of passing information from one to the other sequentially (Fig. 3). More advanced RNNs introduce internal memory to each network to store past information, and make adjustments to the predictions accordingly.

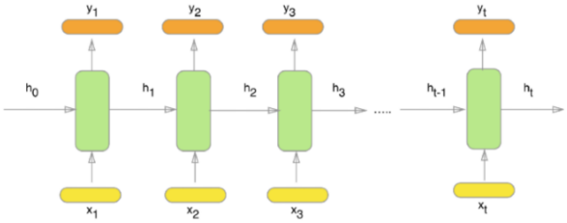
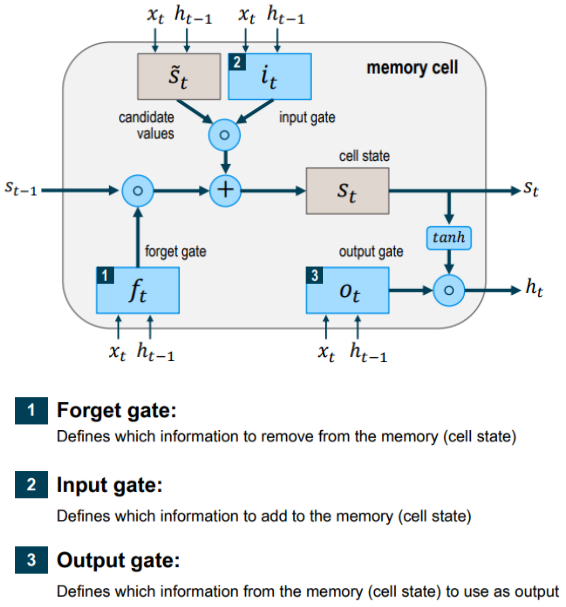


Figure 3: Basic Architecture of an RNN. x_i denote inputs, y_i denote outputs, and h_i denote information exchanged.

However, basic RNNs lack control structures. The lack of control structures make it possible for the long-term components in each memory unit to cause either an exponential growth or decay in the norm of gradients during training, which result in exploding and vanishing gradient problems respectively (Ergen & Kozat, 2017).

To address these inherent issues in RNNs, Long-Short Term Memory (LSTM) networks, a form of artificial RNN, was selected as the final machine learning model. LSTM networks are a state-of-the-art technique for learning from sequential data which are inherently suitable for time series predictions due to its ability to deal with the vanishing gradient problem. Such behaviour renders it insensitive to time gap length in time series data. This enables LSTMs to capture long term dependencies on the data, giving it a considerable advantage in terms of prediction accuracy over regular RNNs. Such patterns can be easily observed in real life trends of parking occupancy data, where occupancy rates can fluctuate due to holidays and festivities which are repeated annually.

LSTM networks are composed of an input layer, one or more hidden layers, and an output layer. Every hidden layer contains multiple LSTM units, which also contains a Constant Error Carousel (CEC) unit (Fig. 4). These CEC units enable LSTM neural networks to discover and memorize the importance of events that happen thousands of discrete time steps ago (Schmidhuber, 2015).



1 **Forget gate:**
Defines which information to remove from the memory (cell state)

2 **Input gate:**
Defines which information to add to the memory (cell state)

3 **Output gate:**
Defines which information from the memory (cell state) to use as output

Figure 4: Architecture of LSTM memory cell

LSTM networks have been demonstrated to be effective at addressing similar problems in various other fields, such as language modelling and speech recognition (Sundermeyer, Schluter, & Ney, 2012), financial market predictions (Fischer & Krauss, 2017), and even as artificial intelligence for games (OpenAL et. al., 2019).

Model Implementation

LSTMs can be difficult to train, requiring large amounts of data and processing power (Culurciello, 2018). Thus, it was decided to train the model on a subset of carparks near serangoon's MRT for a limited time window for demonstration purposes, with the assumption that a model that performs well in a subset of carparks in Singapore will likely generalize well to other carparks.

Data Collection

To verify that the proposed LSTM model can deliver sufficiently accurate predictions of carpark occupancy data, carpark occupancy of seven carparks operated by the Housing Development Board (HDB) around Serangoon MRT in Singapore was obtained for the years 2018 and 2019. This location was chosen due to the high foot and vehicular traffic in the area as well as the relatively high number of carparks within a small radius of approximately 500 meters around the station due to the station being a major interchange with a large shopping mall attached to it.

Initially, historical carpark occupancy was to be obtained from DataMall, a public suite of data operated by the Land Transport Authority (LTA) of Singapore (Land Transport Authority, 2021). This dataset can provide real-time carpark occupancy of many carparks operated by around Singapore. However, this dataset was unable to provide past occupancy data that is necessary for training an initial model. Thus, the occupancy data was instead obtained from data.gov.sg, an open dataset maintained by the Government of Singapore (Open Government Products, 2021) that provides historical occupancy data in addition to real-time streaming data for a more limited subset of carparks operated by HDB.

As it was hypothesized that other factors, such as weather of the area around the carparks also influences the carpark occupancy, the weather conditions (including air temperature, wind speed and direction, humidity, and rainfall) of a weather station located close by, Kim Chuan Weather Station, was also obtained from data.gov.sg.

Although it is hypothesized that there are various other factors that may influence carpark occupancy (such as traffic conditions around the area and whether any significant events were occurring in the area), there was significant difficulty in obtaining data reflecting such conditions and were excluded from the experimental model.

The dataset used for the experimentation in this project is deemed to be comparatively easy to obtain and can also be obtained easily if the model were to be used for the applications described in earlier sections of this paper. This is because existing APIs maintained by the Government of Singapore can already provide this data in real time.

Feature Engineering

Feature engineering aims to create data that contains the most complete and influential features required to make accurate predictions on the carpark occupancy.

Using the dplyr package in R, empty values were replaced with average weekly values for rainfall, temperature, and humidity as there exist days without values for these attributes. Since the problem to be solved is a time series prediction, additional details on the date and time of the entry can influence the predictions significantly. Thus, three boolean indicator variables reflecting if the day was a public holiday, whether it was during a work hour or whether it was a workday were also added to the dataset. This was determined by referring to official work routines and public holidays dates provided by governmental websites. These indicators were verified to be important as the number of vacancies vary significantly in each of these cases and will contribute significantly to the prediction. For instance, during non-working hours, the mean number of carpark lot vacancies was on average, 30% higher than during working hours (Fig. 5).

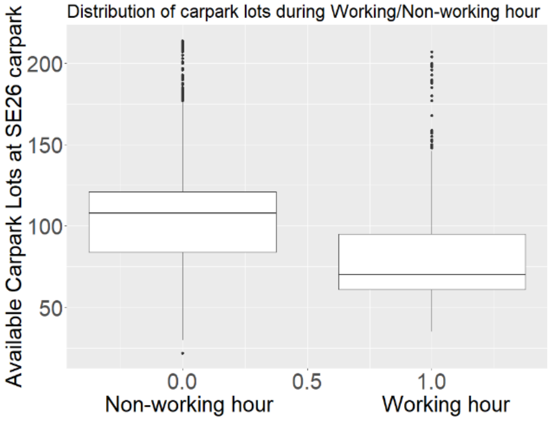


Figure 5: Distribution of carpark lots during working hours and non-working hours

Model Structure

The LSTM network was implemented using the Keras library (Keras, 2021) and its architecture is inspired by Britz (Britz, 2015). The implemented LSTM model was trained on a sliding window size of 3 sequential time frames. The model architecture consists of 2 layers, one input LSTM layer with 107 neurons and 150 memory cells, and one output layer with 7 neurons, each corresponding to a single carpark for a total of 119857 parameters including parameters stored in the memory cells of the LSTM. The activation function used for the LSTM layer is the ReLU function. The final model was trained with a batch size of 250 for 50 epochs, with mean-squared-error (MSE) as the loss function using the Adaptive Moment Estimation method (ADAM).

Results and Discussion

The feature transformed dataset had a train-test split of 80% and 20% respectively. Unlike the usual case where the data is shuffled to eliminate sequential biases, care was taken not to shuffle the input dataset as providing the model with future data to predict past data would go against the model's intended purpose.

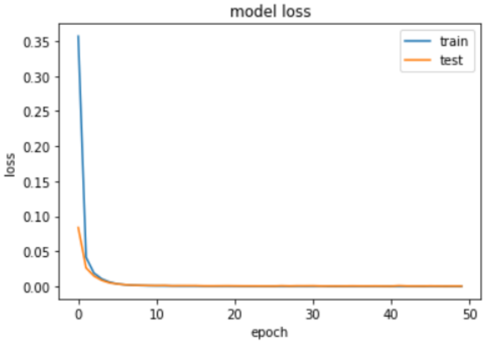


Figure 6: Model MSE against training epoch for the LSTM model

The model weights were observed to converge quickly, achieving a validation MSE lower than 0.001 across all 7 carparks after just 15 training epochs (Fig. 6). The final validation MSE after 50 training epochs was 0.00003 with a validation mean absolute error (MAE) of approximately 0.01 across all 7 carparks, which is more than sufficient for the purposes of the project as carpark vacancies are integral values and need only be accurate to the integer.

To investigate whether the model can be used to predict carpark vacancies far into the future without additional training, the same dataset was used to train another model of the same architecture, but with a 50% split between training and validation datasets instead. The validation MSE of this version of the model was, as expected, worse than the initial model, with MSE values of up to 33 and MAE values of up to 3 per carpark. Thus, it was concluded that the model's predictions tend to drift after longer periods of time, which is hypothesized to be due to the non-stationary nature of carpark occupancy rates that can fluctuate with residency rates and the popularity of the shopping mall over time.

Further Improvements

As shown earlier, the LSTM model produces predictions with a higher MSE and MAE when predicting carpark occupancy rates over a longer period of time. While classical methods, such as log differencing, so exist that can make the test data stationary, such methods are unsuitable for our application as they are more suitable for trend prediction rather than value prediction, which is our key requirement (Press, 2018). Hence, to minimize error from non-stationary trends in data, it is essential for the model to be trained on real-time streaming data. The current model implementation is trained using ADAM, which is unsuitable for online training of the model as it requires a retraining of the entire network for every new data point. To address this limitation, Stochastic Gradient Descent (SGD) could be used for the online training of LSTM network instead, as outlined in (Press, 2018). As SGD allows the model to be improved as and when each new training example arrives, the model can be effectively trained on real-time data without abandoning the result of previous training iterations. Given a large enough dataset, which is available by extracting more past data, and a sufficiently small learning rate, SGD is able to approximate batch gradient descent which was used in the initial implemented model.

Alternative Models

To evaluate the need for using LSTM networks, which is a rather complicated model, to predict future carpark occupancy, the performance of simpler Non-Neural Network models for time series forecasting was evaluated. It was determined that these alternative, simpler models were inferior to LSTM networks at predicting future carpark occupancy, which validates the hypothesis that the temporal dimension plays a significant role in the prediction.

Gradient Boosted Trees

Gradient boosted trees rely on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error (Fig. 7). While gradient boosted trees are considerably simpler and easier to use compared to LSTMs, they lack the ability to model the sequential nature of data, and thus requires the problem to be transformed from a sequence prediction task to a supervised learning problem.

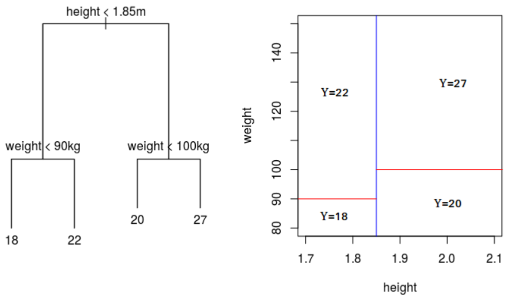


Figure 7: Depiction of a Tree Regressor

The XGBoost model, a gradient boosted tree, segments features into ranges and learns different response values for each segment by adjusting for the respective ranges of each feature. It is incapable of memorizing or considering the sequence of the data, thus ignoring potential patterns that may be identified in the time data sequence, which was hypothesized to be essential for prediction.

Furthermore, LSTM models, being a neural network, was found to experience greater improvements in accuracy when trained with more data, but not for the XGBoost model. When provided with the same data, the MSE of the predictions by the XGBoost model was 9100, faring much worse than the LSTM. The XGBoost model is thus less robust than the LSTM, predicting well only with a small dataset, requiring more effort on the part of the user to preprocess data when deploying it. It can only give meaningful predictions when the training dataset spans a duration of up to five months or 3000 entries (Fig. 8).

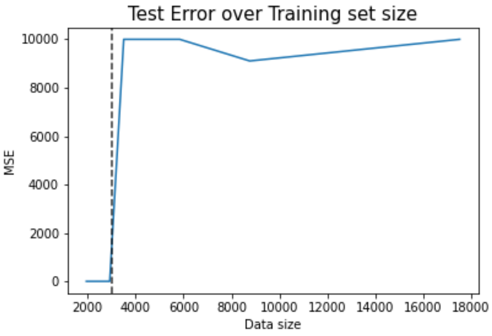


Figure 8: Graph of MSE against data size of the XGBoost model

In addition, the XGBoost model, even at its optimal data size, only achieved an MSE of 0.09, significantly worse than the LSTM's MSE by a factor of 10^{-2}. Hence, LSTM was demonstrated to be more effective than alternative models.

Ethical Implications

In theory, dynamic pricing helps distribute congestion much like the ERP system in Singapore encourages drivers to drive at certain times or take alternative routes, which helps alleviate heavy traffic congestion. Dynamic pricing encourages drivers to park at certain times for cheaper rates and eases overcrowding at certain carparks, which benefits everyone. However, this application can result in inequality among drivers as those incentivized to not visit during crowded times are drivers who are less financially capable and are more affected by pricing whereas the rich benefit more from dynamic pricing as they are inelastic to changes in pricing and would benefit the most from less crowding of carparks. This results in inequality as this application clearly favors the rich over the poor.

Because collected data is never perfect, it may have missing values. There is an ethical implication when it comes to filling in missing attributes, as the process carries bias, which affects the output of the model. This may be unfair to consumers as they are treated differently compared to if we had a perfect dataset free of bias.

Due to the nature of black-box models, it is difficult to explain or interpret how the model derived its conclusion, which can result in unfavorable conclusions that are biased towards certain groups. For example, our model could discriminate the religious people more as they congregate at certain timings out of necessity to go for church services. There have been real-world examples, such as Apple's credit card service giving men higher credit card limits even though it should have been blind to gender (Pena, 2020). This bias was eventually found to have arisen from other variables. Such unintended drawbacks are a possible ethical implication as it is difficult to 'teach' the model morals that give us the desired outputs.

Conclusion

In conclusion, this project has taught the team about the possibilities of using machine learning to ease the lives of the average Singaporean. One important takeaway is that machine learning should only be used when there is a significant value in its predictions, that is, machine learning should not be done for the sake of doing machine learning.

Through this project, it was also discovered that industry applications of machine learning involve several different many steps, from problem analysis and application design, to data collection, model selection, implementation, and integration. Each step necessitates a unique set of technical skills. This project also exposed the group to many different types of models outside the scope of what is taught in an undergraduate machine learning course, which exposed us to the complexity of machine learning algorithms. This knowledge will definitely be useful for future endeavors.

References

Britz, D. 2015. Recurrent Neural Network Tutorial, Part 4 – Implementing a GRU/LSTM RNN with Python and Theano. Retrieved March 21, 2021, from wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/

Culurciello, E. 2018. The Fall of RNN / LSTM. Retrieved March 21, 2021, from towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0

Ergen, T. Kozat, S. S. 2017. Efficient Online Learning Algorithms Based on LSTM Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 29(8): 3772-3783. doi.org/10.1109/TNNLS.2017.2741598

Fischer, T., Krauss, C. 2017. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* 270(2): 654-669. doi.org/10.1016/j.ejor.2017.11.054

Kee, D. M. 2020. An assessment of the viability of the smart parking system: The case of a smart city initiative in Malaysia. *Global Business and Organizational Excellence* 39(5): 26-34. doi.org/10.1002/joe.22013

Keras. 2021. Keras: Deep Learning for Python. Retrieved March 21, 2021, from github.com/keras-team/keras

Kostadinov, S. 2017. How Recurrent Neural Networks work. Retrieved March 21, 2021 from towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7

Land Transport Authority. 2021. Land Transport DataMall. Retrieved March 21, 2021 from datamall.lta.gov.sg

NCCS. 2017. Singapore's Emission Profile. Retrieved March 21, 2021 from nccs.gov.sg/singapores-climate-action/singapore-emissions-profile/

OpenAI et. al. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. Retrieved March 21, 2021, from cdn.openai.com/dota-2.pdf

Open Government Products. 2021. Data.gov.sg. Retrieved March 21, 2021 from data.gov.sg/

Pena, A. S. 2020. Bias in multimodal AI: testbed for fair automatic recruitment. In Proceedings of of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020: 760-761. doi.org/10.1145/3382507.3421165

Press, J. 2018. LSTM Online Training and Prediction: Non-Stationary Real Time Data Stream Forecasting. *Detroit: Wayne State University, Department of Computer Science.*

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61: 85-117. doi.org/10.1016/j.neunet.2014.09.003

Shekhar, A. 2019. Understanding The Recurrent Neural Network. Retrieved March 21, 2021 from afteracademy.com/blog/understanding-the-recurrent-neural-network

Sundermeyer, M., Schluter, R., Ney, H. 2012. LSTM Neural Networks for Language Modeling. *Interspeech*.

Roles and Contributions

Chua Zong Wei: Researching and identifying potential models to address problem statement, LSTM model hyperparameter training and tuning, applying data clustering and analysis techniques for data preprocessing.

Marcus Duigan Xing Yu: Research and writing of ethical implications, research the possible applications of project to real world scenarios and how they benefit society overall. Organized meetings and deadlines for the team.

Munnamgi Harsha Vardhan: Researched on the background of the issues related to car parking as well as to write up on how the results could be used in real world applications. Did up the problem analysis for this project.

Teh Nian Fei: Initial data collection, finalization of LSTM model architecture and hyperparameter tuning, creating graphs to visualize model performance, and overall structure, outline and report editing

Teo Hoe Keat: Initial data collection, provide an exact real-world application of the model, including mock-up images of apps and how the implemented model can be integrated into the apps.

Zou Runzhong: Feature engineering on the multiple datasets to produce two aggregated datasets with significant feature columns. Run experiments on the XGBoost model for comparison with the LSTM model.