Deep Learning to Handle Congestion in Vehicle Routing Problem: A Review

D Satyananda¹² and A Abdullah¹
¹School of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor Bahru, Johor, Malaysia.
²Department of Mathematics, Universitas Negeri Malang, Jl. Semarang 5, Malang 65145, Indonesia Malaysia.

Email: satyananda@graduate.utm.my

Abstract. This paper reviews the implementation design of Deep Learning in Vehicle Routing Problem. Congestion and traffic condition are usually avoided in Vehicle Routing Problem due to its modeling complexity, and even the benchmark datasets only cover essential conditions. In the real situation, the traffic condition is varied, and congestion is the worst part. To model the real life, the delivery route must consider these situations. The vehicle needs information on traffic prediction in future time to avoid congestion. The prediction needs historical traffic data, which is very large. Deep Learning can handle the enormous size and extract data features to infer the prediction.

1. Background

Vehicle Routing Problem (VRP) is the problem of designing optimal delivery or collection routes from one or more depots to some customers, subject to one or more constraints [1]. There is a central depot storing customer's goods, customers and their demand for the goods, and a set of vehicles with a specific capacity to distribute the goods from the depot to the customers.

During the distribution process, each customer is only visited once by a vehicle, and the vehicle is required to go back to the starting depot to form a close loop, a Hamiltonian cycle. VRP is commonly modeled using a graph, as in Figure 1. D is the depot node, and other nodes are customers. Distance between nodes is written on each edge that connects a node to another node. Q is the list of customer's demands.

![Figure 1. Typical graph model for VRP case](image)

The objective of VRP is to find the least delivery cost (least total distance). Many VRP references...
consider that VRP has known and fixed cost or travel time [2] and mostly aimed to minimize total distance without considering some other important aspects [3]. By assuming constant vehicle speed, then minimizing distance equals minimizing travel time. In fact, the speed varies during the time, and the traffic sometimes unpredictable. This condition makes VRP modeling sometimes unrealistic.

Many efforts have made the vehicle speed varies, or even worse, the traffic is congested. The traffic congestion is divided into predictable or Recurrent Congestion (RC) and unpredictable or Non-Recurrent Congestion (NRC) [4,5]. NRC is caused by a specific characteristic of the area (junction, intersection, railroad crossing, public places) and common rush hour. Traffic dynamicity can raise extra distribution costs or travel time. The congestion in a specific point can propagate along a fixed direction and a fixed path to a neighboring point, impacting traffic on a regional scale [5].

If the drivers recognize the congestion previously, they can change the distribution route by taking alternative paths by which the delivery mission is completed without entering other congestions. Due to its emergency nature, it is unavoidable that the new route could be longer than the previous most optimal route. Hence, involving the traffic factor is essential in the distribution route arrangement and during the tour. Accounting for the traffic condition has considerable potential for cost savings [4].

Many researchers discussed congestion problems. They analyzed existing traffic data to evaluate traffic conditions, predict congestion, and avoid congestion. The size of the traffic data itself is massive because it is recorded over time. Deep Learning (DL) can provide features and information from big-size data because it can automatically extract features from large-scale raw data. It has been successfully applied in various domains, such as computer vision and speech recognition [6].

The transportation network consists of thousands of links with changing traffic conditions over time. DL can address high-dimension features of traffic networks through distributed representations; thereby, it is very promising in learning high-dimensional features with tremendous data [7]. Its powerful capabilities would be an excellent system for traffic modeling and predicting traffic conditions in the VRP distribution process.

This paper discusses DL implementation to detect congestion and its avoidance in VRP cases. Previous research excludes the involvement of current and future traffic conditions in VRP. To continue the scholarship, the present study was designed to (i) find a valid and effective DL model for handling various spatial and temporal traffic data to predict congestion and (ii) incorporate traffic and congestion factor in VRP route arrangement. The study shed lights on a better VRP route arrangement by using traffic information to avoid congestion.

2. Introduction to Deep Learning

DL is a subfield of Machine Learning (ML) and Artificial Intelligence (AI) subsequently, to automatically learn from the concepts and knowledge without being explicitly programmed [8]. What differentiates ML and DL is that DL employs a collection of ML algorithms. DL has a multi-layer architecture to discover complex structures and patterns. Different layers capture features from different perspectives to form a multi-level abstraction, compared to classic ML models (such as SVM and ANN), which only have a “shallow” architecture to capture features [6]. The typical architecture of DL is depicted in Figure 2. There is an input layer with many neurons, several to thousand connected hidden layers, and an output layer. The hidden layers transform the input layer states into the output layer's expected inferences by capturing the high-level abstractions. The number of input and output values depends on the extracted features, and it can be different for each model.

After the feature extraction phase, a set of the most relevant features is selected based on some criteria to simplify the modeling and enhance the generalization capability. The traffic model is constructed using the most informative and non-redundant features. The model consumes training data to find the exact feature and evaluate against testing data. This process is iterative and repeated until a specific performance is achieved, involving parameter tuning to get the most accurate prediction.
3. Congestion detection using Deep Learning

Congestion detection is only a part of traffic forecasting, and it is related to travel-time forecasting. Lana et al. [9] reviewed road traffic forecasting and described a list of traffic forecasting aspects. They found that only one highly relevant literature about congestion prediction during 2014-2016, from a total of 63 works of literature about traffic predictions. It is mostly about travel time and traffic flow.

It also happens in the DL. The research mostly analyzes traffic conditions and less for congestion detection. In traffic modeling, historical data from many sources are processed to extract its features. The features are correlated to specific traffic conditions. The condition is influenced by spatial dimension (such as a road segment’s neighbor) and temporal dimension (peak and non-peak, weekend and weekday, special and regular event). These spatiotemporal dimensions determine the evolution and play an essential role in predicting future congestion states.

The traffic data sources can come from many sources, as discussed in [6,9-11]. Among those, data from online map providers are preferred because they are publicly accessible, there is no special equipment needed, have vast auxiliary information, and provide current data. Real-time data is more suitable for predicting or detecting congestion while traversing the route.

The primary data for traffic analysis is the road network, GPS trace, and GPS Trajectory. The road network describes the spatial topology of the transportation infrastructure. GPS trajectory is a sequence of discrete spatial points containing traffic information related to human mobility [12].

The speed data for 24 hours is divided into several time ranges. More time ranges are more realistic, but it needs more computational effort. The average speed of all GPS points in the road segment for all time ranges defined is used to obtain traffic information in the road segment. The traffic information is converted into a congestion matrix. There are some notations to measure congestion conditions [13]. This matrix considers a spatiotemporal condition: a condition on a specific place and time. Each row states a specific point or road segment, and each column states a specific time slot. DL model uses this matrix to predict congestion conditions for some times after $t$. Figure 3 is a general form of the congestion condition matrix, with $C_r^T$ denotes congestion condition in a road segment $r$ in a specific time range $t$.

![Figure 2. Typical architecture of deep learning](image)

![Figure 3. Congestion condition matrix](image)
The additional input of the model depends on the problem detail. To detect congestion propagation in neighboring points, it needs data of connected segments. Auxiliary information such as weather conditions, road closure, peak hours, public events, specific places on the segment (market, railroad crossing, junction, and others) is useful for predicting traffic conditions based on non-recurrent events. Prediction of recurrent congestion is easier than non-recurrent because there is a typical pattern on the data, such as weekend and weekday, peak and non-peak hour. Historical data is not adequate to give enough information to predict future non-recurrent traffic conditions.

Some prevalent DL models include Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Stacked Autoencoder (SAE) [6]. LSTM is an improvement of the RNN; it uses memory cells (special hidden neurons) to remember input for a long time, which RNN cannot handle. LSTM is designed for long sequential data and automatically determines the optimal time lags for prediction [6]. Hence it is suitable for sequential traffic data. It can recognize a correlation in traffic input, and it is widely employed in traffic information prediction, which has promising results [14]. There are variants of LSTM used for traffic detection problems, for example, Seq2Seq model that has encoder-decoder built from LSTM [11], a fusion of LSTM and GRU (Gated Recurrent Unit) to better capture the spatiotemporal features of the lanes’ travel speed [14], bidirectional LSTM that considering backward propagation to increase prediction accuracy [15], and integration of ARIMA and LSTM to capture linear and non-linear features at a time [16].

4. Deep Learning for Vehicle Routing Problem

Incorporating DL in VRP undertakes two main processes:

i. Building traffic detection model using Deep Learning

Some steps to build the model include:

a. Collecting GPS data to depict the roadway traffic condition. The GPS data may come from highway authority or public-domain data published. The invalid, redundant, and erroneous data are removed. The data are aggregated into a spatiotemporal matrix (see Fig. 2). Elements of the matrix are vehicle speed or the value of congestion condition notation. Other than GPS trajectory, the input is also geographic information (road segment length, number of lanes, neighboring point, specific point characteristic, and usual or specific events).

b. Learning process using Deep Learning. Predicting spatiotemporal congestion evolution pattern uses LSTM model family.

c. Validating prediction results using testing data. The testing data use some part of the dataset or using real-time data gathered from online map data. It takes some parameter tuning to minimize validation error. Validation uses Root Mean Square Error (RMSE), Root Mean Square Logarithm Error (RMSLE), and Mean Absolute Error (MAE) test.

d. Visualizing the current state and future prediction in a digital map.

ii. Arranging VRP delivery route

The process consists of:

a. Arranging a VRP route using information from Deep Learning. The output of this model is a value that expressing traffic conditions in some future time. The value is converted into average vehicle speed to count travel time between customers. Route arrangement uses Sequential Insertion algorithm and Variable Neighborhood Descent (VND) optimization.

b. During the travel, the traffic situation is also monitored using real data from the online map provider. In case there is a situation leading to traffic congestion, a congestion avoidance plan must be carried out. Alternative routes are searched using the shortest path search algorithm. Dijkstra algorithm can be exploited to indicate a route without congestion to maximize the road traffic flow [17]. The selection must consider traffic conditions around the congestion point to avoid the vehicle is trapped in another congestion and avoiding too many deviations from the original route.

c. Visualizing the new information in a digital map.
5. Conclusion
This paper discusses the implementation of variable speed and congestion factors in VRP cases. The idea reaches a realistic route arrangement considering the traffic situation. This system implements Deep Learning to analyze traffic situations, predict future traffic conditions, and detect if a specific road segment is congested. The feature extraction model uses LSTM, which is famous for handling sequential information. The result of the Deep Learning system is used as a consideration in arranging the VRP delivery route. The vehicle speed and congestion factor involvement could give a more realistic route arrangement and save delivery cost and time. Future research is encouraged to find sufficient requirement to model traffic congestion, find the most optimal DL model, and investigate reinforcement learning to predict non-recurrent congestion.

References
[1] Laporte G 2009 Transp. Sci. 43 408–416
[2] Malandraki C and Daskin M S 1992 26 185–200
[3] Alvarez P, Lerga I, Serrano A, and Faulin J 2017 Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10268 39–50.
[4] Kok A L, Hans E W, and Schutten J M J 2012 Comput. Oper. Res. 39 910–8
[5] An S, Yang H, Wang J, Cui N, and Cui J 2016 Inf. Sci. (Ny) 373 515–26
[6] Liu Z, Li Z, Wu K, and Li M 2018 IEEE Netw. 32 40–6
[7] Ma X, Yu H, Wang Y, and Wang Y 2015 PLoS One 10 1-17
[8] Dargan S, Kumar M, Ayyagari M R, and Kumar G 2020 Arch. Comput. Methods Eng. 27 1071–92
[9] Lana I, Del Ser J, Velez M, and Vlahogianni E I 2018 IEEE Intelligent Transportation Systems Magazine 10 93–109
[10] Wang S, Djahel S, and McManis J 2016 IEEE Vehicular Networking Conference, VNC 2016 196–203
[11] Liao B, Mellonraith D, Zhang J, Chen T, Wu C, Yang S et al. 2018 Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining 537-46
[12] Wang Y, Cao J, Li W, Gu T, and Shi W 2017 Pervasive Mob. Comput. 41 470–83
[13] Afrin T and Yodo N 2020 Sustainability (Switzerland) 12 1-23
[14] Gu Y, Lu W, Qin L, Li M, and Shao Z 2019 Transp. Res. Part C Emerg. Technol. 106 1–16
[15] Essien A, Petrounias I, Sampaio P, and Sampaio S 2020 World Wide Web
[16] Lu S, Zhang Q, Chen G, and Seng D 2020 Alexandria Eng. J. 60 87–94
[17] Rocha Filho G P, Meneguette R I, Torres Neto J R, Valejo A, Weigang L, Ueyama J et al. 2020 Ad Hoc Networks 107 1-10