
Bouziane Brik and Adlen Ksentini

**Abstract**

Collision detection and avoidance between vehicles is one of the key services envisioned in the Internet of Vehicles. Such services are usually deployed at the multi-access edge computing (MEC) infrastructure to ensure low-latency communication and thus guarantee real-time reactions to avoid collisions between vehicles. In order to maximize the coverage of the road and ensure that all vehicles are connected to an optimal MEC host (in terms of geographical location), the collision avoidance application needs to be instantiated on all the MEC hosts. This may add a burden on the computing resources available at the latter. In this article, we propose an AI-empowered framework that aims to optimize the computing resources at the MEC hosts. Our framework uses deep learning to predict the vehicle density to be served by a MEC host and derive the exact computing resources required by the collision detection application to run optimally. We evaluate the proposed framework using a real dataset representing vehicle mobility in a big city. Obtained results show the accuracy of our prediction model, and hence the efficiency of our resources assignment framework to exactly deduce the optimal computing resources needed by each instance of the application.

**Introduction**

In the vision of 5G and the Internet of Vehicles (IoV), connected vehicles and autonomous driving are among the key envisioned applications; they impact not only mobile operators as well as car manufacturers business, but also everyday life [1]. Connected vehicles and autonomous driving involve several components, such as sensors, actuators, and applications, which need to coordinate in order to achieve the envisioned autonomy of vehicles. Among the critical services toward autonomous driving is the collision detection/avoidance system. It consists of continuously collecting data from vehicles and using these data to predict collisions and communicate alerts or send commands to vehicles in order to avoid collisions with other vehicles. The collision detection/avoidance service comprises an application that runs at the vehicles and collects data, such as GPS coordinates, speed, and acceleration, and a remote application hosted at the cloud infrastructure that runs a collision detection algorithm. The latter may send control commands to the vehicles, when deemed appropriate, such as reduce speed, change direction, or brake. One of the main requirements to run the collision avoidance service, in 5G and beyond, is to dispose of a low-latency connection between the client and server sides of the application, that is, between the vehicles and the remote application sitting in the infrastructure. To ensure low-latency communication, it is envisioned to deploy several instances of the collision application, for instance, at the road intersections and close to base stations using multi-access edge computing (MEC) [2]. This ensures that each vehicle is connected to the closed application instance, which guarantees low-latency communication. MEC consists of deploying computation capability (hosts) close to the end users, for instance, in the vicinity of base stations [3]. Thus, all data can be treated locally without involving the remote cloud server, hence reducing latency and the traffic to carry throughout the network. MEC hosts are distributed all over the network, constituting a distributed and low-latency computation resource for delay-sensitive applications like vehicle collision detection/avoidance applications. Indeed, the latter need low-latency communication with remote vehicles since the control commands, such as brake or reduce speed, need to be received by the vehicles in near real time to react to any threat and avoid collisions. As mentioned earlier, one pertinent solution is to locate several instances of the collision detection application at the MEC, hence reducing the end-to-end communication latency. The different instances can be used to cover all the road intersections; each vehicle shall always be in contact (connected) with an instance of the application. However, duplicating the number of instances of the application may add a burden on the MEC host computing resources, including CPU usage, memory, storage, and so on. It is worth noting that MEC host resources are limited compared to central cloud servers. In this context, it is vital for MEC operators (generally network operators) to optimize the MEC resource usage, particularly when considering that 5G will rely on, among others, MEC to support services that require low latency such as data caching and virtual/augmented reality (VR/AR). Hence, ensuring efficient sharing of the MEC computing resources is critical.
In this article, we propose a novel framework that relies on deep learning (DL) to predict vehicles’ mobility and accordingly assign computing resources to the vehicle collision detection application instances, aimed at better optimizing the overall MEC resources and ensuring optimal functioning of the service (i.e., guaranteeing low response time). The proposed framework predicts using DL, and more precisely long short-term memory (LSTM), the mobility of the vehicles, and, according to their positions in the network, it derives the needed MEC resources: on one hand, increasing MEC resources for the collision avoidance application instances that serve a high number of vehicles; on the other hand, reducing these resources for the instances that cover a lower number of connected vehicles. The contributions of this article are as follows:

- Define an AI-empowered framework that relies on European Telecommunications Standards Institute (ETSI) MEC to optimally deploy the vehicle collision detection and avoidance service.
- Train and build a DL module based on LSTM, using a real dataset to predict the vehicles’ mobility.
- Devise a resource assignment algorithm that runs at the MEC orchestrator (MEO) and considers the mobility prediction of vehicles when assigning computing resources to the running application instances.

The rest of the article is organized as follows. The following section presents the state of the art on the usage of MEC to support vehicle collision detection and avoidance systems. Following that, we describe our proposed framework in terms of mobility prediction model as well as resource assignment algorithm. We then devote a section to the performance evaluation of the proposed framework evaluation and the obtained results analysis. Finally, conclusions are drawn.

### Related Work

Employing the MEC system to optimize the collision detection and avoidance service in IoV has rarely been explored in the literature. Only a few works exist. These works usually start by collecting vehicles’ mobility information in real-time before detecting collisions and alerting the concerned vehicles. In [4], the authors address one of the relevant classes of automotive services, called extended virtual sensing (EVS) services, which, among others, shows the benefit of using edge computing resources in the context of autonomous driving. In particular, the authors focus on the performance of the collision avoidance system. In fact, these works are based on simple schemes to estimate the next trajectories of mobile users, for instance, by only determining the distances between vehicles in real-time [4] or reading the status of the blinking lights [6], while in this article, we explore the usage of DL to predict vehicles mobility and use this prediction to improve the management of MEC computing resources.

### Proposed Framework

#### Architecture

As stated earlier, we assume that the collision avoidance application is duplicated (i.e., several instances) and deployed on all the MEC hosts of a network operator. Allowing coverage, with low-latency access, of a large geographical area. We recall that a MEC host includes a virtualization platform that runs applications’ instances in the form of virtual machines (VMs) or containers. Each MEC host has a computing capacity depending on the used hardware, which is limited compared to a centralized cloud server. Another drawback is the collision detection scheme, which depends mainly on the performance of the collision avoidance system. In fact, these works are based on simple schemes to estimate the next trajectories of mobile users, for instance, by only determining the distances between vehicles in real-time [4] or reading the status of the blinking lights [6], while in this article, we explore the usage of DL to predict vehicles mobility and use this prediction to improve the management of MEC computing resources.

It is vital for MEC operators (generally the network operator) to optimize MEC resource usage, particularly when considering that 5G will rely, among others, on MEC to support services that require low-latency such as data caching, Virtual and Augmented Reality, and so on. Hence, ensuring an efficient share of the MEC computing resources is critical.

In particular, the next vehicles’ trajectories are predicted by projecting the next vehicles’ positions onto curved or straight segments. This projection depends on the status of the blinking lights.

- Collision Avoidance Strategy (CAS) notifies the vehicles potentially involved in a collision regarding the action needed to avoid it.
- In [7], the collision avoidance application for vehicular networks is extended to benefit vulnerable users (e.g., pedestrians and cyclists) equipped with smartphones. The authors proposed a MEC-based collision avoidance system. Through a Basic Safety Message (BSM), this system periodically collects users’ information such as position, acceleration, speed, and direction in order to estimate users’ trajectories and avoid collision between them.

Although the above works addressed the collision avoidance issue between vehicles through the MEC system, they did not consider the MEC resources usage given the limited capacity of MEC hosts compared to centralized cloud servers. Another drawback is the collision detection scheme, which depends mainly on the performance of the collision avoidance system. In fact, these works are based on simple schemes to estimate the next trajectories of mobile users, for instance, by only determining the distances between vehicles in real-time [4] or reading the status of the blinking lights [6], while in this article, we explore the usage of DL to predict vehicles mobility and use this prediction to improve the management of MEC computing resources.

In [7], the collision avoidance application for vehicular networks is extended to benefit vulnerable users (e.g., pedestrians and cyclists) equipped with smartphones. The authors proposed a MEC-based collision avoidance system. Through a Basic Safety Message (BSM), this system periodically collects users’ information such as position, acceleration, speed, and direction in order to estimate users’ trajectories and avoid collision between them.

Although the above works addressed the collision avoidance issue between vehicles through the MEC system, they did not consider the MEC resources usage given the limited capacity of MEC hosts compared to centralized cloud servers. Another drawback is the collision detection scheme, which depends mainly on the performance of the collision avoidance system. In fact, these works are based on simple schemes to estimate the next trajectories of mobile users, for instance, by only determining the distances between vehicles in real-time [4] or reading the status of the blinking lights [6], while in this article, we explore the usage of DL to predict vehicles mobility and use this prediction to improve the management of MEC computing resources.
Later we describe how we can obtain these values.

To recall, in MEC, the MEO handles the life cycle management (LCM) and orchestrates the computing resources of the MEapps. It is in charge of deploying the MEapps on top of the virtualized platform at the MEC host. The MEapps are described using an application descriptor (AppD), which includes configuration information, such as the application image, and the computing resources needed by the application. The MEO is then in charge of requesting the CPU resources and updating the request if deemed appropriate by scaling up or down the needed CPU, that is, increase or decrease the number of vCPUs assigned to an MEapp. For more details on ETSI MEC architecture, interfaces, and components, readers may refer to [8].

In this work, our objective is to derive at the MEO, for each MEapp instance, the needed number of vCPUs by finding a trade-off between optimizing the MEC host computing resources and ensuring low response time to optimally run the service. Intuitively, one solution would be that each time a vehicle (or a batch of vehicles) has moved from one MEC host to another, the resource management algorithm running at the MEO computes the needed vCPUs for each application instance it manages. However, this solution requires continuous tracking of vehicles’ mobility at the network layer, which is difficult to enforce. Accordingly, we propose to leverage the above-mentioned solution with a mobility prediction model using LSTM to anticipate the update rate of resources (vCPUs) needed by each MEapp instance. The aim is to not only optimize the MEC resources, but also ensure that an MEapp instance performs optimally (low response time) considering the number of vehicles connected to it.

The proposed algorithm runs at the MEO, as shown in Fig. 1. It is composed of two modules. The first one, namely the mobility prediction module, takes as inputs the vehicle GPS coordinates obtained from the collision avoidance MEapps (via Mp1 interface [8]) and predicts, through the LSTM module, the next position of the vehicles. The second module, namely LCM decision, uses the predicted next vehicle positions to derive the necessary computing resources (vCPUs) of each MEapp instance, which allows optimal running of the collision avoidance service. The MEO enforces the LCM decisions via the MEC edge platform manager (MEPM), which, according to ETSI, is in charge of updating the CPUS resources of the running MEapps through the MEC host/edge virtualization infrastructure manager (MEP/VIM) [8].

The global algorithm runs as follows:

- **Epoch Loop**
  1. Mobility prediction: Receive a batch of GPS coordinates from the MEapp.
  2. Mobility prediction: Predict the next location of vehicles for epoch t + 1, and hence the number of users to be connected for each MEapp instance.
  3. LCM: Run a decision algorithm that (1) takes as inputs the predicted number of served vehicle per MEapp for t + 1, (2) extracts the number of users to be connected for each MEapp instance.
necessary computing resources according to the predicted number, and (3) changes the configuration of the MEapp instances through the MEPM if deemed appropriate.

In the next sections, we detail the mobility and resources prediction as well as LCM modules.

**VEHICLES MOBILITY PREDICTION MODEL**

To predict the vehicles’ mobility, we use deep recurrent neural networks (RNNs) with an LSTM algorithm. The main procedures are the dataset, designing the neural network, training the network, and testing the network. The three first steps are described in the following subsections, while the last step is described in the performance evaluation section.

**Taxi Mobility Data:** We use real and publicly available taxi trace data, which is composed of 464,019 records and gathered over 30 days in San Francisco, California [9]. This dataset was collected in May 2008 and contains mobility traces in terms of GPS coordinates of approximately 500 taxis. Each taxi is equipped with a GPS module and sends periodically, each 10 s, its location (timestamp, ID, geo-coordinates) to a central server.

It is worth noting that we consider the San Francisco map as \((n \times m)\) grid cells, and we translate each taxi’s GPS coordinates to a cell ID. The dataset is used to provide actual taxis’ GPS coordinates to our mobility prediction model in order to predict the next ones, and hence the vehicles’ next location, which is used as input by the resource assignment algorithm to update or not the assigned computing resources to the running MEapp instances. We use this dataset rather than simulated GPS coordinates to validate our algorithm under real mobility traces.

**Design of the Taxi Mobility Prediction Model:**

An RNN with an LSTM algorithm is well suited to classify, process, and predict time series, given time lags of unknown duration [10]. In fact, RNN with LSTM is capable of learning long-term dependencies between input data by using an internal memory to remember past data memory. This makes it suitable for our problem to predict the next location of vehicles (Cell IDs) based on the past one.

Figure 2 illustrates our LSTM-based vehicle mobility prediction model. The prediction process comprises three main steps. The input vehicle’s trajectory is first processed by a fully connected input layer with 56 neurons. Each vehicle’s position (Cell ID) is mapped to a 56-dimensional feature tensor. Then the resulting sequence is sent to a deep RNN composed of three stacked LSTM layers, each with 56 neurons.

Each LSTM layer considers the previous LSTM layer’s output as input and feeds its output to the next LSTM layer. Finally, a fully connected output layer with 45 neurons maps the output of the last LSTM layer to the cell ID, corresponding to the predicted vehicle’s cell ID of the next time step, \(i_{t+1}\). We argue for the usage of 45 neurons due to the fact that we divided the San Francisco map into 45 cells.

Finally, the training of the model phase aims to minimize the distance between the predicted and real location of vehicles (Cell ID). Hence, we choose the mean squared error (MSE) as the loss function and adopt the Stochastic Gradient Descent algorithm to update the neural network parameters [11].

**LCM Resources Assignment**

As indicated earlier, the LCM module runs the decision algorithm, which may request resources update for the running instances of the collision avoidance MEapp. Let assume \(v_t\), \(loc_t\), \(u_t\) as vectors that represent the vCPU used by an MEapp, the location of an MEapp (MEC server ID), and the number of vehicles connected to an MEapp, respectively. The index of the vector corresponds to the collision avoidance MEapp instance number. We note by \(v_t, u_t, t\) the value of the vectors at epoch \(t\). At the initial epoch \((t = 0)\), we note \(v_0 = (c_0, c_0, .., c_0)\), and \(u_0 = (0, .., 0)\), where \(c_0\) corresponds to the initial configuration of the MEapp (i.e., number of vCPUs), and \(u_0\) is the initial number of connected vehicles to the collision avoidance application. The decision algorithm is detailed in Algorithm 1, where \(N\) is the number of instances, \(\text{Change}(\cdot)\) is a function that gives the necessary vCPUs to optimally handle \(X\) users. \(C(X)\) is an integer value between 1 and \(M\). Note that \(C(X)\) can be derived by benchmarking an instance of the collision avoidance application, which can be obtained by simulation or using a real deployment. The \(C(X)\) function aims to indicate the number of users that can be handled by one instance while ensuring the computing latency is very low, hence reducing the response time.

First, the decision algorithm considers as input the predicted number of vehicles to be served by each MEapp instance for the next epoch. Second, it verifies if the current number of vCPUs used by the MEapp is not optimal (more resources are needed or overusage of the resources). If so, an update of the resources is requested. Finally, for all the MEapp instances that need to be updated, a request is sent to the MEPM.

**Performance Evaluation**

We divide the performance analysis of the proposed framework into two parts: the performance and accuracy of the mobility prediction module, and the LCM resource assignment performance.
For the first part, we used the Tensorflow engine to implement our RNN with the LSTM-based prediction model [12]. In addition, we compared our learning model with two other learning algorithms:

- RNN with gated recurrent unit (GRU) model, which is similar to LSTM, but has fewer parameters than LSTM, as it lacks an output gate [13]
- Convolutional neural network (CNN) model, a deep learning algorithm designed to process arrays of data that can also be applied for time series forecasting problems [14]

Table 1 presents the considered parameters to compare the three algorithms. For the sake of fairness in terms of comparison, the three algorithms share the same activation function and optimizer (stochastic gradient descent).

For the second part, we simulated a distributed MEC system that covers the city of San Francisco. The city map has been divided into \((n \times m)\) grid cells. We assume that the grid cells are covered by six MEC hosts, and each host is in charge of covering a set of cells. Furthermore, we assume that each collision avoidance instance needs one vCPU to manage two vehicles. It should be noted that this number can be fixed by benchmarking the application in real deployment or by simulation. Finally, we suppose that each MEC host has 15 vCPUs, which means that an MEapp instance may get a maximum of 15 vCPUs; thus, it can manage up to 30 vehicles. We then compare our resources assignment scheme to a static scheme that assigns a fixed number of vCPUs (10) to each MEapp instance, whatever the vehicle density (i.e., the number of vehicles to be served by a MEapp instance). We focus on two main metrics to validate our scheme: overloaded MEapp instances (i.e., the demand exceeds server capacity in terms of vCPU resources) and overprovisioned MEapp instances (i.e., less than 20 percent of an MEapp’s vCPU resources are used); which corresponds to a high response time of the application instance and a non-optimal usage of the MEC host computing resources, respectively.

**Mobility Prediction Evaluation**

Figure 3a compares the considered learning algorithms’ performance in terms of MSE on the test dataset, aiming at validating the performance of our prediction model using unseen data, that is, data that the models have not seen before. We remark that our LSTM-based model minimizes the MSE compared to the other algorithms even when we increase the number of test samples.

![Figure 3a](image-url) 

**Figure 3a.** Performance comparison between learning algorithms: a) mean squared error on the test dataset; b) real and predicted cell ID values.

![Figure 3b](image-url) 

**Figure 3b.** Real and predicted cell ID values.

**Table 1.** Implementation parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td></td>
</tr>
<tr>
<td>Number of records</td>
<td>464,019</td>
</tr>
<tr>
<td>Number of cars</td>
<td>500</td>
</tr>
<tr>
<td>Collection time period</td>
<td>May 2008</td>
</tr>
<tr>
<td>Percentage of training set</td>
<td>80% of the dataset</td>
</tr>
<tr>
<td>Percentage of test set</td>
<td>20% of the dataset</td>
</tr>
<tr>
<td>Deep learning</td>
<td></td>
</tr>
<tr>
<td>Deep learning tool</td>
<td>Tensorflow</td>
</tr>
<tr>
<td>LSTM timestamp window</td>
<td>60</td>
</tr>
<tr>
<td>Activation functions</td>
<td>Relu (hidden layers)  Softmax (output layer)</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Stochastic gradient descent (SGD)</td>
</tr>
<tr>
<td>Loss function</td>
<td>Mean squared error</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.01</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.2</td>
</tr>
<tr>
<td>Batch size</td>
<td>50 samples</td>
</tr>
<tr>
<td>Number of epoch</td>
<td>[20, 60] epochs</td>
</tr>
<tr>
<td>MEC</td>
<td></td>
</tr>
<tr>
<td>Number of MEC hosts</td>
<td>6 hosts</td>
</tr>
<tr>
<td>Number of vCPUs</td>
<td>15 vCPUs</td>
</tr>
</tbody>
</table>

**TABLE 1.** Implementation parameters.
ID values of 10 test samples. Mostly, we notice that the predicted cell IDs using the LSTM-based model are similar to the real cell ID values. However, RNN with GRU and CNN algorithms fail to predict the correct cell ID for some test samples (e.g., test samples ID = 2, 3, 8, and 9). These results confirm Fig. 3a’s results, that is, the efficiency and the accuracy of the LSTM-based model in predicting cars’ cell IDs.

**Resource Assignment Evaluation**

Figures 4a and 4b represent the vehicle density under the coverage of each MEC host (and hence the number of vehicles to be served by an MEapp instantiated at the MEC host) during a rush hour (from 8 a.m. to 9 a.m.), and the number of overloaded MEapp instances during that hour, respectively. The number of overloaded MEapp instances is a critical metric for the service-level performance as an overloaded MEapp means high response time, which degrades the performance and may lead to safety issues. From Fig. 4a, we observe that the vehicle density is high; most of the MEapp has more than 20 vehicles to serve, which requires more than 10 vCPUs. We remark in Fig. 4b that our scheme keeps the number of overloaded MEapp instances very low compared to the static scheme. We argue this by the fact that the LSTM-based model can predict the density of vehicles in each group of cells and hence can anticipate the needed vCPU resources of the collision avoidance instances (i.e., MEapp instances), which allows adapting to the vehicle density to be served. However, by fixing the vCPU value to be used by each MEapp instance, the static scheme results in a high number of overloaded MEapp instances. This is mainly due to the fact that the demand (vehicle density) exceeds the MEapp instances’ capacity (i.e., 20 vehicles), which can clearly be observed in Fig. 4a.

Figures 5a and 5b illustrate the vehicle density during low traffic hours (from 3 p.m. to 4 p.m.) and the number of overprovisioned MEapp instances during that hour, respectively. From Fig. 5a, we observe that the vehicle density is low, and most of the MEC host covers fewer than 10 vehicles, which requires only 5 vCPUs. We notice from Fig. 5b that again, our scheme minimizes the number of overprovisioned MEapp instances compared to the static scheme. In fact, anticipating the required resources also helps to reduce the number of overprovisioned MEapp instances, as the LCM computes exactly the needed number of vCPUs to use with the current vehicle density to serve. However, in the static scheme, several MEapps are overprovisioned. Only 5 of 10 vCPUs are used to manage the current vehicle density, which leaves 5 unused that cannot be assigned to another MEapp instance in the MEC host.

To summarize, we can deduce that our resource assignment algorithm’s performance depends mainly on the accuracy of the LSTM-based prediction model. The generated results show the efficiency of our LSTM-based model to estimate the number of vehicles at each cell, which in turn improves the efficiency of our resource assignment algorithm in forecasting exactly the needed resources by each MEapp instance, hence improving the overall resource of the MEC host. Contrariwise, the static scheme uses the same computing resources, whatever the density of vehicles, to serve for all the MEapp instances, which leads to degrading both the MEC computing resource usage and the service performance.

**Conclusion**

In this work, we introduce an AI-empowered framework that aimed to improve the management of MEC resources (mainly computing) when deploying vehicle collision detection and avoidance service in IoV. The framework aims to find a trade-off between improving the overall usage of MEC resources and guaranteeing that each instance of the deployed vehicle collision detection and avoidance application is assigned enough resources to optimally run (i.e., low response time). To find this trade-off, our framework uses RNN with LSTM to predict vehicle density at each cell and then computes the exact computing resources needed by each instance of the application.

We built our RNN with LSTM-based model and evaluated our framework using a real dataset of vehicle mobility in a big city. The obtained
results show the accuracy of our prediction model in estimating the exact number of vehicles at each group of cells and the efficiency of the resource allocation resource to optimize both the overall MEC resources and the application performance.

One of our future work directions is to perform position prediction in a distributed way, using distributed learning, which will allow the vehicle to make the prediction locally.

Acknowledgments

This work has been partially supported by the European Union’s H2020 MonB5G (grant no. 871780) project and the ANR Chist-era Leading Edge project.

References


Biographies

BOUZIANE BRIK received his Ph.D degree from Laghouat and La Rochelle Universities, France, in 2017. He is currently working as an associate professor at Burgundy University and DRIVE Laboratory. He has been working on network slicing in the context of the H2020 European project on 5G. His research interests also include the Internet of Things (IoT), IoT in industrial systems, smart grid, and vehicular networks.

ADLEN KSENTINI is an IEEE Communications Society Distinguished Lecturer. He obtained his Ph.D. degree in computer science from the University of Cergy-Pontoise in 2003. Since March 2016, he is a professor in the Communication Systems Department of EURCOM. He has been working on several EU projects on 5G, network slicing, and MEC.