 DOI 10.1109/JIOT.2021.3098973

Automatic Distributed Deep Learning Using Resource-constrained Edge Devices

Alberto Gutierrez-Torre, Kiyana Bahadori, Shuja-ur-Rehman Baig, Waheed Iqbal, Tullio Vardanega, *Member, IEEE*, Josep Lluís Berral, *Member, IEEE*, David Carrera, *Member, IEEE*

***Abstract*—Processing data generated at high volume and speed from the Internet of Things, smart cities, domotic, intelligent surveillance, and e-healthcare systems require efficient data processing and analytics services at the Edge to reduce the latency and response time of the applications. The Fog Computing Edge infrastructure consists of devices with limited computing, memory, and bandwidth resources, which challenge the construction of predictive analytics solutions that require resource-intensive tasks for training machine learning models. In this work, we focus on the development of predictive analytics for urban traffic. Our solution is based on deep learning techniques localized in the Edge, where computing devices have very limited computational resources. We present an innovative method for efficiently training of Gated Recurrent-Units (GRUs) across available resource-constrained CPU and GPU Edge devices. Our solution employs distributed GRU model learning and dynamically stops the training process to utilize the low-power and resource-constrained Edge devices while ensuring good estimation accuracy effectively. The proposed solution was extensively evaluated using low-powered ARM-based devices, including Raspberry Pi v3 and the low-powered GPU-enabled device NVIDIA Jetson Nano, and also compared them with Single-CPU Intel Xeon machines. For the evaluation experiments, we used real-world Floating Car Data. The experiments show that the proposed solution delivers excellent prediction accuracy and computational performance on the Edge when compared with the baseline methods.**

***Index Terms*—Internet of Things (IoT), Edge Computing, Resource Management, Big Data, Analytics, Cloud Computing, Fog Computing**

I. Introduction

THE Internet of Things (IoT) is attracting significant interest from both academia and industry. The potential benefit of applying IoT paradigms to Smart Cities and Health Care service scenarios suggest to design new architectures for infrastructure, platforms and services. The issue with more traditional approaches rises from the inherent limitations in connectivity and computing power of Edge devices and dynamic networks. Those IoT architectures are usually composed of real-time sensor-based monitoring systems and actuators running in different locations, connected to data aggregation applications or data-warehouses through dynamic networks (such as 5G, Wi-Fi or wired Internet).

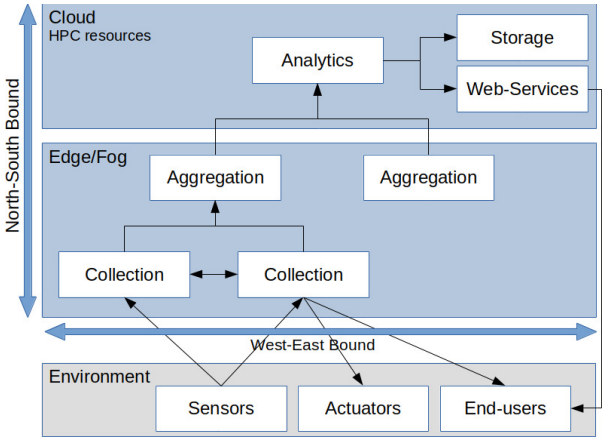


Fig. 1. Edge-Cloud Aggregation Schema, with environment actors (Sensors, Actuators and Users).

The main feature of the Cloud is to provide extremely scalable resources to service applications from a remote datacenter. In contrast, emerging scenarios like the IoT, smart cities, domotic, intelligent surveillance, and e-healthcare usually require proximity and quick reaction time while generating massive amounts of data transmitted to the analytics applications. Fog computing is more attractive for such demand [1]. Fog computing takes the computation to the Edge, moving data processing close to the sources, and reducing data to synthesized volumes to be transmitted north-bound to the Cloud, as shown in Figure 1. Additionally, when the Edge services depend only on local data, the service can be provided without using Cloud services. Several fields can benefit from this kind of architecture, specifically Oil & Gas [2], power grid systems [3], smart cities, smart industries, and IoT applications [4]. In these environments, local analytics are required as they need a low latency QoS [5]. Moreover, back-haul connectivity might fail [6] as the network might not be as reliable as wanted due to extreme conditions. Given the importance of exploiting the data at the Edge level, considerable research effort was devoted to establish a common framework to cope systematically and effectively with the restrictions proposed by this kind of environment [7].

The compute-intensive nature of the training of Machine Learning (ML) models has so far caused that all the processing is done in Cloud data centers. This typical strategy, to push the data to the cloud and then training the ML models, has the advantage of using powerful computing machines.

Alberto Gutierrez-Torre, Josep Lluís Berral and David Carrera are with the Barcelona Supercomputing Center and Polytechnic University of Catalonia. e-mail: {alberto.gutierrez, josep.berral, david.carrera}@bsc.es

Kiyana bahadori and Tullio Vardanega are with the University of Padova. e-mail: bahadorikiana@gmail.com, tullio.vardanega@unipd.it

Shuja-ur-Rehman Baig and Waheed Iqbal are with the University of Punjab. e-mail: {shuja, waheed.iqbal}@pucit.edu.pk

However, this strategy has several drawbacks: it adds a cost of additional network dependency, increases latency, and moves the processing away from the data producers. In contrast, using limited computing power available at Fog nodes is interesting for training ML models efficiently. Recent work shows the importance of training an ML model on Edge infrastructure. For example, Plastiras et al. [5] show the importance of doing the computation for training Deep Learning models on Fog nodes for Computer Vision tasks like object detection. The authors remark the importance of privacy, performance, latency, and power efficiency on this kind of application, which can be a perfect fit for Edge and Fog Computing.

In this work, we investigate the use of Fog devices to train deep learning models by distributing the training on available devices at the Fog environment intelligently. Our proposed solution takes benefit of already active devices aggregating or collecting data instead of employing additional Cloud resources. This also reduces network communication and protects services from network disruptions by keeping them autonomous on the edge. This concept extends the work by Perez et al. [6], where local models in the Edge level can be independent of global ones. This way the architecture is resilient against back-haul network interruptions. The model synchronization can be delayed to the moment when network is available.

We present a system to automatically distribute the time-consuming task of training deep learning models on a Fog computing network consisting of low-powered and resource-constrained computing devices. The proposed approach is based on Federated Learning (FL), which leverages the work of McMahan et al. [8] and Bonawitz et al. [9]. The proposed solution automates part of the Deep Learning process for selecting appropriate parameters for the model to reduce the training time while maintaining the model accuracy for validation data set. We extensively evaluate the proposed system using a road traffic analytics scenario designed for city-wide traffic modeling and prediction running on the Fog computing paradigm. The proposed methodology can make use of any kind of Neural Network (NN) by distributing the training on Fog devices. In particular, we use Gated Recurrent Unit (GRU) neural networks to model the traffic behavior to produce short/medium-term traffic predictions following the FL principles. Our evaluation investigates different data aggregation levels, different levels of data processing parallelism, time requirements for achieving suitable accuracy levels for models, and suitability for real-time applications in the Edge. Our evaluation is based on real traffic logs from one week of Floating Car Data (FCD) in Barcelona. The data was provided by one of the largest road-assistance companies in Spain and comprises thousands of vehicles. This approach is tested in a Smart City setting, however the same approach can be used for other fields like Oil & Gas, where distributed learning is required or desirable. Moreover, one appealing domain to apply our proposed solution is the healthcare industry where patient data is collected through IoT devices and required to process locally without sending remote locations for privacy and security concerns.

The experimental results show that predictive analytics requiring complex ML mechanisms like GRUs can be performed cost-effectively on the Fog nodes without using expensive Cloud resources. Additionally, compared to prediction methods previously used in other studies, we show that GRUs achieve good accuracy results with constrained training time in comparison using state-of-the-art methods (i.e., Conditional Restricted Boltzmann Machines (CRBMs)). Even though the modeling process is split to reduce training time, the distributed model shows a stable behavior when modifying training hyper-parameters. The research contributions of this paper are as follows:

- A system for distributed modeling for city-wide applications using the Fog computing paradigm for predictive analytics using low-powered and resource-constrained devices.
- A mechanism to automate run-time decisions for stopping training processes when accuracy levels are reliable for deep neural networks.
- Evaluation and comparison of time required to model the deep neural network using the proposed solution on Fog (low-power and resource-constrained) vs. Cloud (high-performance) environments.
- A comparative analysis of resource usage vs. accuracy on training models for real FCD compared with existing baseline methods.

This paper is organized as follows: Section II reviews the background and motivation. Section III illustrates the proposed architecture. Section IV describes our approach methodology. Section V shows the evaluation, and Section VI provides concluding remarks and future challenges.

II. State of the Art

A. Background and Motivation

The Cloud has been widely used to address the emerging challenges of big data analysis in many smart city ecosystems such as smart houses, smart lighting, and video surveillance [10], [11], [12]. However, IoT scenarios usually require low latency between sensors/actuators and usually there are scarce computing resources. These restrictions avoiding unnecessary north-south bound communication of data that can be processed on the Edge or intermediate nodes [6]. Location awareness is also a must in several Smart Cities IoT architectures providing immediate in-place services. As IoT services in Smart Cities are being increasingly used, Cloud services alone can hardly satisfy the mentioned requirements of this ecosystem.

Fog computing, the paradigm combining the Edge and Cloud capabilities, can handle the significant data treatment, including acquisition, aggregation, analytics and preprocessing, while reducing transportation and storage, even balancing computation power among intermediate nodes [13].

In addition, transforming this data into actionable knowledge and adapting to changing dynamics of modern cities, requires *intelligent* modeling techniques not only accurate but adaptive. ML techniques enable *smartness* in Smart Cities by modeling, predicting and extracting useful information from

collected data, through advanced statistics and artificial intelligence algorithms. Deep Learning, a ML subfield based on multi-layer neuronal networks, is becoming an important tool to city-modeling challenges across many areas such as forecasting [14], self-driving research [15], image processing [16], [17], or object recognition [18], useful to manage public services, detect hazardous scenarios or to guide emergency services among others.

However, the increasing amount of data to be processed, along with the computational demands of sufficiently-accurate neural network algorithms, have led to bigger computational and memory resource requirements. Accelerating neural networks training to competitive accuracy within a sufficiently short time is a major challenge that may lead to increase computational demands. Seeking solutions that assure scalable and efficient learning has given rise to the notion of "distributed ML". Federated Learning (FL) [8] is a promising solution when both data and resources are scattered along in the architecture, with the added challenge of the near impossibility of having all data in the same place, and the cost of constantly offloading computation to the Cloud.

FL aims at distributing the data or, as in the case of Edge computing, keeping the data near where it is produced [8], [9]. This solution can be understood as allowing the Edge devices, the clients, to produce a predictive model with their own local data, and then coordinate with a central node, the server, for model merging. In particular this is interesting in the contexts where data privacy is an issue as in the work of McMahan et al. [8], as the only data exchanged between the data producers and the central server are the weights, i.e. the configuration, of the neural network. On the other hand, there have been efforts like in the work of Hu et al. [19] that focus on having a model that works properly on both sides, client and server. Moreover, it has been proved Stochastic Gradient Descent (SGD) converges in this scenario [20], proving the suitability of Neural Networks for this particular task. This approach brings about properties that are desirable for Edge Computing architectures, like the ability to keep on working without network connectivity when the system fails [6].

Even though the methodology per se is already available, there still is a knowledge gap regarding the actual applicability of FL on a Fog architecture using low-powered devices. This work aims to fill this gap applying the methodologies described in the following sections.

B. Related Work

The exponential growth of the IoT, caused by the opportunity of leveraging smart devices in generalized enterprise settings, motivates the quest for novel approaches to develop deep learning system that can scale to very large models and large data set. However, training to competitive accuracy within a sufficiently short time span, for large and complex networks together with huge data sets is especially challenging in Edge/Fog nodes at the present state of the art.

A significant amount of effort and research has been devoted to tackling the challenge of training huge data sets through building large models with more parameters and parallelization or distribution methods based on the Cloud computing infrastructure. For example, Google implemented a distributed framework for training neural networks over Central Processing Unit (CPU) based on the DistBelief framework [21], [22] which makes use of both model parallelism, and data parallelism. This model has also proved useful for computer vision problems, achieving state-of-the-art performance on a computer vision benchmark with 14 millions of images.

To scale up the training phase of learning, researchers utilize accelerators such as a single or cluster of Graphics Processing Units (GPUs) [23], [24]. Recently, Facebook [25] announced achieving 90% scaling efficiency in training visual recognition model, using data parallelism combined with the use of GPUs.

K. Hong et al. [26] proposed a fog-based opportunistic spatio-temporal event processing system to meet the latency requirement. Their system predicts future query regions for moving consumers, and starts the event processing early to make timely information available when consumers reaches the future locations. Yu et al. [27] proposed a Deep Reinforcement Learning based system that is able to share execution of tasks in Edge nodes taking into account the battery, quality of service and other details.

Works such as Marchisio et al. [28] study how to perform ML inference in ultra-low powered devices, and review the usage of NNs with this kind of device. This approach minimizes both power usage and hardware costs. Sudharsan et al. [29] proposed a methodology to train a kind of Convolutional Neural Network (CNN) and then adapt it to run in different MicroController Units (MCUs) to do prediction. Their approach reduces the size of the trained network to the 10% of the original. In the same direction, TinyML [30] enables training a NN with TensorFlow and then convert it to it can be run using TensorFlow Lite on ultra-low power MCUs. Neither of these approaches handle training on the device, but other approaches like Neuro.ZERO [31] enable training on the device by means of hardware acceleration. However, FL has yet to be covered on this kind of setup with MCUs, so that it enables to train different models and average the model configuration among nodes.

To the best of our knowledge, there currently is no evaluation of this kind of problem with FL using Recurrent Neural Networks (RNNs) with server-class hardware and low-powered devices. Moreover, mechanisms are needed to stop training as soon as a reliable-enough model is obtained. We believe that FL distributed learning can be highly beneficial for data analytics over scenarios like smart cities.

III. Architecture: Floating Car Data Processing Over Edge

This section presents our proposed architecture for processing Floating Car Data (FCD) using Edge computing infrastructure. We explain the Edge computing network, FCD, and data processing pipeline in the following subsections.

A. Edge Computing Networks

Edge computing networks are based on architectures where sensors collect data from nearby cars, users, and equipment and send them to the computing nodes within proximity.

Such nodes are low-powered with limited resources to perform complex analytics; therefore, the data is pushed to the remote Cloud for processing using sophisticated and powerful hardware. The "Fog" is that part of the architecture embracing Edge nodes receiving data from sensors, Intermediate nodes performing intermediate data aggregation, and Cloud APIs receiving data to be processed and stored, extending the Cloud paradigm [32]. Figure 2 shows a Fog infrastructure, with near-data nodes on the Edge, intermediate nodes with medium power to pre-process aggregate or localized data, and the Cloud.

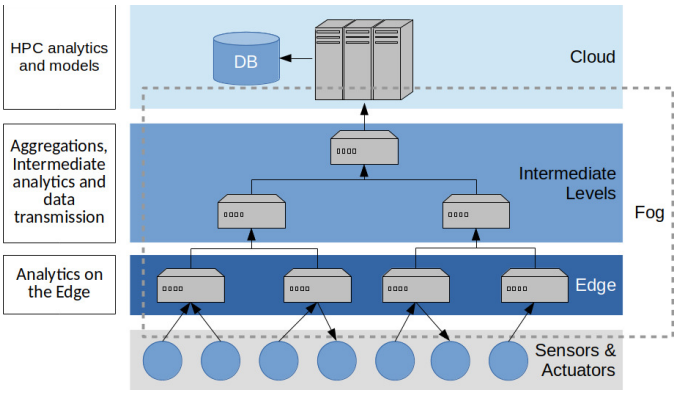


Fig. 2. Schema of the Fog Infrastructure, from Edge to Cloud.

Current devices on the Edge are specially designed to consume low power, produce low throughput, and offer low capabilities, such as Raspberry Pi and NVIDIA Jetson. These devices with Edge computing have been recently considered a good solution for smart city image processing challenges [33], showing that industry and public administration are interested in adopting the approach of using low-powered and resource-constrained devices for real scenarios. NVIDIA Jetson [34] is a low-power and small form factor computer similar to Raspberry Pi (ARM processor). It is a Linux-enabled machine that is equipped with an embedded NVIDIA low-power GPU and the CUDA framework, which can be used to train Deep Learning calculations. However, a single Jetson device is not sufficient to perform the deep learning training independently for a large data set.

In this paper, we propose using the low-power and resource-constrained devices to train the deep learning model at the Edge, close to the users and data, by distributing the training on multiple devices and enabling the Edge efficient analytics. In our system, sensors collect data and transmit to the Edge nodes, and analytics are performed on the Edge nodes instead of offloading the complex analytics tasks to the Cloud.

B. Floating Car Data

FCD represents geo-localized timestamped data of moving vehicles, collected and analyzed for various applications, including smart cities, traffic engineering, and traffic management. Typically, FCD is received through antennas deployed in the town representing a large urban zone, a localized neighborhood, a street, or a street segment, depending on which granularity is required for the specific application. That data is provided to the Edge analytics indicating the received timestamp for each vehicle transmission and its speed. Data like vehicle position, e.g. Global Positioning System (GPS), is not provided for privacy and security reasons; only the Edge node position is provided.

The FCD arrives asynchronously to our Edge nodes and is aggregated periodically into summaries of traffic information, i.e., the average speed of vehicles surrounding the node (in Km/hour) and vehicles' count, considering that vehicles will be reported once for each aggregation window time. Considering the aggregation of a 1-minute interval as lower bound, we aggregate the incoming data into data entries containing latitude, longitude, number of cars, speed average and timestamp. Before performing the analytics, the Edge nodes independently collect and aggregate the FCD into a specific time interval.

The 1-minute aggregation data is only a base for larger aggregations, as traffic time-series can be aggregated from minutes to hours to days because of its periodic pattern in time. While large aggregations can be easily predicted due to this periodicity, smaller aggregations can be more challenging. For the validation experiments, in Section V, we test different levels of time aggregation varying from 5 minutes to 1-hour intervals for training analytic models.

In this work, we used a week-worth of real FCD from the city of Barcelona, Spain, provided by one of the largest road-assistance companies in the country.

C. Data Analytics Pipelines

Whenever FCD is detected through antenna sensors, it is transmitted to the nearest Edge node. The data aggregation using a specific time interval is performed at the Edge node. For each aggregation, the timestamp is added to the FCD record for building a time-series data set. The FCD time-series data set is used to model the traffic behavior for forecasting and analytics purposes. Figure 3 shows the FCD collection, modeling, and forecasting pipeline. In our system, we performed distributed model training, explained in Section IV-C, using low-powered Edge nodes.

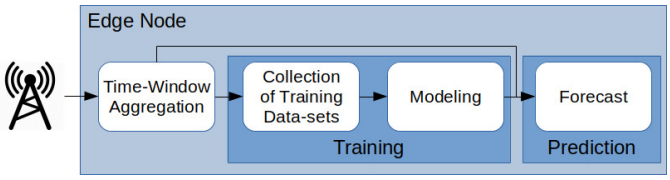


Fig. 3. Pipeline of Time-Window Aggregation, Learning and Prediction.

For a global-scale prediction, the aggregated data and created models on the Edge can be pushed to the Cloud for storage and further analysis. Moreover, the aggregated data from individual Edge nodes can be passed to intermediate nodes for training a generalized model, as depicted in Figure 2. However, in previous works [6], we observed that local models fit local scenarios better than general models in the Cloud, avoiding intense communication interruption problems.

IV. Methodology

This section presents our proposed methodology for FCD time-series forecasting, automation for the training process, and distributed model learning on the Edge.

A. Traffic Forecasting

We can see the FCD time-series data set as a matrix of size number of time window elements times input features. The forecasting problem targets the prediction of two variables: the number of cars and the average speed of the following time step $(t + 1)$ using the previous d elements from the time window, where d is our *delay* or memory window. As the time window is a whole aggregation period, the goal is to predict the next period of traffic information. We used GRU networks [35] to train the forecasting model. Given the capabilities of the GRUs, it is possible to forecast far from $t + 1$, as GRUs are shown to be capable of medium-term forecasting in many scenarios. GRUs are *generative*, and can generate predictions by using their last prediction and status as input/memory for the next prediction. In our problem, we are predicting from $t + 1$ up to $t + N$, where N is the size of testing data set in the experiments (approximately 1 day in the following experiments).

B. Training Process Automation

ML model training with good accuracy is controlled by a "Training vs. Validation" process. In NNs, this process is used to decide when to stop the iterative process. Training data is divided into two batches: "training" and "validation" data sets. The longer NNs trains with the whole training data set (each period is called an epoch), the more fitted the model is expected to be, but only to the training data, which can lead to *over-fit*. To mitigate this risk, the validation data set is predicted at each *epoch*, allowing to check how the model behaves with "non-training" data. While the error in training data decreases at each epoch, error in validation data decreases until the point of *over-fitting* and increases from there, as Figure 5 illustrates. That point is considered the "bouncing point", and data-scientists would manually stop iterating at that point. But for non-stable data, as we face with FCD, validation can differ enough from training data on certain occasions, and those expected behaviors may not be encountered in during training. Hence, the time at which the process should stop stop iterating must be decided automatically.

To detect the bouncing point to stop the deep networks' training process for minimizing the training time while achieving good accuracy, we propose Algorithm 1. The algorithm shows the technique for fixing p (point of bounce) dynamically, using the error on training and validation, previously smoothing both sequences to facilitate treating *wavy* sequences, using a Locally Estimated Scatterplot Smoothing (LOESS) curve fitting method [36]. Among other algorithms with the same objective, LOESS was selected for its simplicity and speed. It is well fitted for low-powered devices and, for our use case, it achieved good results for low computational cost.

Algorithm 1 Detecting GRU training cutting-point (epoch) p from training and validation error

```
Result: p point of bounce/intersect/convergence/minimum,
        prioritizing error on validation over error on training
smooth_tr, smooth_val ← loess_smooth(error_tr, error_val)
if exists_bounce(smooth_val) then
    return bounce_p(smooth_val)
else
    if exists_intersect(smooth_tr, smooth_val) then
        return intersect_p(smooth_tr, smooth_val)
    else
        if exists_bounce(smooth_tr) then
            return bounce_p(smooth_tr)
        else
            if converges(smooth_tr, min_threshold) then
                p ← converging_p(smooth_tr, min_threshold)
                return min(p, minimum_p)
            else
                return minimum_p
            end
        end
    end
end
```

The process of finding p implies running for a given amount of epochs, to find the trend and detect the bouncing, intersection or convergence point. This process can be substituted by more sophisticated methods that can be applied online, although the set of rules we have devised can be used to determine p once for a given amount of data, while retaining p for future models.

C. Distributed Model Training

Computing devices at the Edge are low-powered and very limited in terms of the number of cores and storage. The available processing power is mostly used to receive and transmit data from sensors to Cloud; remaining computational resources can perform aggregation and modeling processes.

In our proposed distributed model training solution, we assume the availability of single CPU/GPU processors on each available Edge device. To distribute training across workers (available Edge devices), we partition the training-validation data set and send it over to the available workers. Each worker creates a model from its subset and validates it. At that point, all sub-models are joined in the initiator Edge node and merged following the FL principles [8]. The resulting aggregated model can either work better for the dilution of noise among sub-models or do worse due to over-fitting each sub-model to its sub-set. For this reason and good practice, the aggregation model is evaluated on the test data set in the initiator Edge node. Figure 4 shows the process of distributed training, merging and evaluation. This process is done in a off-line fashion using the whole data set, but it also could be done receiving a stream of continuous data, retraining the networks for new batches and synchronizing after each epoch.

In our test case, we have split the data evenly between nodes using the data coordinates to split regions. Then inside each

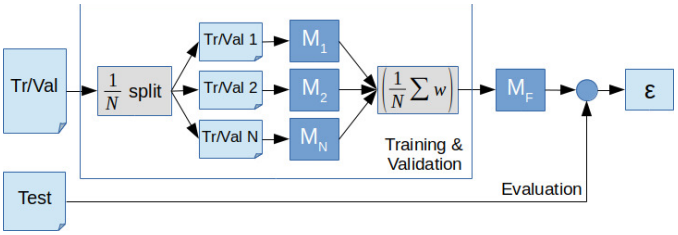


Fig. 4. Distributed modeling technique for training and testing phases. Data split among N workers, creating N models to be merged, creating the final model to be evaluated.

node, the split of training vs. validation is done following a 80% – 20% ratio between training vs. validation, for every subset. Each worker splits its data to train and validate its model. The test set (a 20% of the total data) is kept for the aggregated model for evaluating the final model.

V. Experimental Evaluation

The proposed approach was evaluated with several experiments designed to test the model learning and accuracy on the previously mentioned real FCD set with one week worth of data. We compare two approaches: a single learning model that learns from all the data versus multiple local models that are synchronized in a FL framework. We compare the effectiveness of the different learning model configurations in low-powered and resource-constrained Edge devices. Implementation and evaluation of the proposed solution were performed using TensorFlow and Keras frameworks over R. Notice that R can run on any Linux-enabled device and that the core of the code is built on top of TensorFlow, which is efficiently implemented in C++. The infrastructure to run and measure training times corresponded to a server-class single thread Xeon processor for comparison experiments among different training configurations and two low-powered devices: a Raspberry Pi 3 (ARM processor), and an NVIDIA Jetson Nano (ARM processor + NVIDIA GPU) to cover both CPU and GPU settings.

Our experiments addressed the following evaluation aspects:

1) The effects of training the GRU with a different number of Hidden Units and a different number of epochs, and check the usefulness of determining a stop-point p dynamically using the presented set of rules versus fixing a large enough p *a-priori*.
2) The comparison and trade-off between training epochs vs. hidden units vs. resulting error vs. level of time aggregation.
3) The effects of distributing the training process among N different processors, considering a low range for N matching the dimensions of common low-power devices.
4) The capability of running the presented methods on low-powered devices, i.e. Raspberry Pi v3 and NVIDIA Jetson Nano.

A. Hyper-parameter Identification

We evaluated the capability of the Deep Neural Network (DNN) to learn the target time-series of *Volume (Cars)* and *Average Speed (Speed)* of traffic data using the proposed Algorithm 1 for identifying appropriate epochs. We also ran a grid search-like strategy for determining the number of hidden units and time aggregation levels, i.e., periods in which data is aggregated into a single value for the proposed solution.

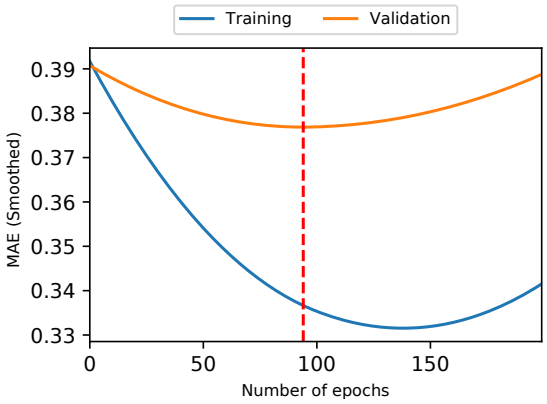


Fig. 5. Zoomed representation of the smoothed Mean Absolute Error (MAE) as used in Algorithm 1. Observe that at 94 epochs the validation data bounced back, selecting it as a training stop-point.

In this experiment, we trained a single model using the entire training data set and our proposed Algorithm 1 to automatically identify the number of epochs with higher accuracy. We evaluated various settings for hidden units and aggregation levels. In the case of 2 hidden units and 20-minute aggregation, our algorithm identified 94 epochs as a bouncing point, as shown in Figure 5. To compare the proposed solution for identifying epochs, we have performed additional experiments and manually tuned the epochs from 10 to 200. Notice that the stop decision is made with the validation data, as doing so with training data could lead to overfitted models.

Figure 6 shows the error distribution for different training epochs over the *Test* data set and the dynamic stop-training point using Algorithm 1. For the volume of cars, learning seems easy as we observed MAE between 1 – 1.3. Predicting the speed of cars becomes more complex as we yield volatile error between 3 – 5.5 in Km/h (the variability of the traffic speed on those data sets is already known from previous works [6]). While most of the training is done on the first few epochs of the different tested NN configurations, identifying automatic stop-training point becomes conservative with respect to the best option, but performing almost as good as the optimal.

The experimental results reported in Figure 6 show the difficulty of establishing a set of rules that match every single training-validation scenario. Selecting the best number of epochs is still an open problem whose solution can be automated with more complex mechanisms. However, for the current scenario, where quick decisions must be made, the presented algorithm becomes an adequate solution. Therefore, from now on, the results shown are the ones using the d value for epochs.

Figure 7 shows the Root Mean Square Error (RMSE) for estimating the number of cars on test data for 2, 4, 8, 16 and 32 hidden units with 5, 10, 15, 20, 30 and 60 minutes aggregation levels. We observed the aggregation yields stable

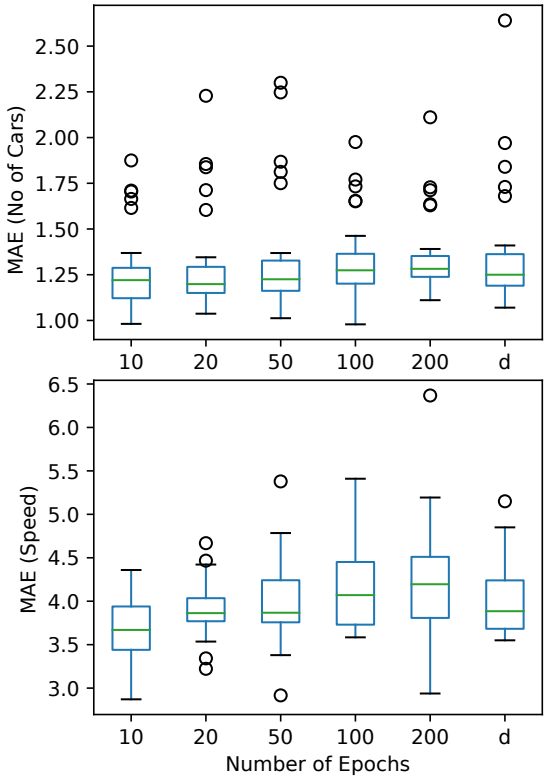


Fig. 6. Comparison of MAE for various static number of epochs with a dynamic number of epochs for estimating the number of cars and average speed. Here d represents a dynamic number of epochs.

behavior until 30 minutes aggregations as the RMSE remains under 2. However, at 60 minutes aggregation level, we observed a significant increase in the estimation error. This is because, with a higher level of data aggregation, the underlying fine-grained details are hidden, and the model cannot learn from data accurately. We observed the effect of changing the number of hidden units does not have any significant effect on accuracy. The average RMSE of estimating the number of cars remains between 1 to 2 except at the aggregation level of 60 minutes.

Figure 7 shows the RMSE for estimating the speed of cars on test data for 2, 4, 8, 16 and 32 hidden units with 5, 10, 15, 20, 30 and 60 minutes aggregation level. We observed the high error for aggregation levels 5 and 60; however, it remains similar for other aggregation levels. We do not observe any noticeable accuracy gain for using a different number of hidden units. The average RMSE of estimating cars' speed remains between 4.5 to 5.5 except aggregation level of 5 and 60 minutes.

This set of experiments allowed us to determine the appropriate level of aggregation and the hidden units to determine to be used in the final model. We computed the average RMSE of speed and number of cars on test data for $(2, 4, 8, 16, 32)$ hidden units with $(5, 10, 15, 20, 30, 60)$ minutes aggregation levels. Figure 8 shows the average RMSE for estimating the speed and number of cars. Each aggregation level has its optimal number of hidden units, meaning that there is no optimal configuration able to deal with all levels of aggregation, a desirable state allowing us to decide the precision of

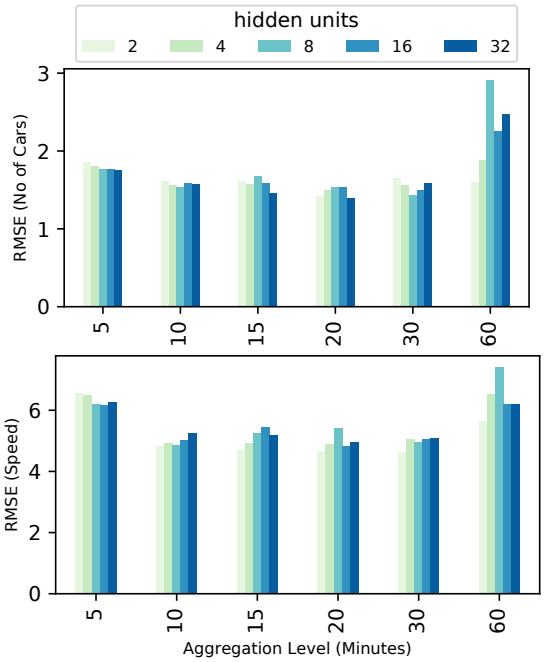


Fig. 7. Error vs. hidden units vs. time aggregation for number of cars and speed estimation with dynamic epoch value d.

the time interval. Determining a time interval where we can trust predictions the most, we observed that the 2 hidden units with 20 minutes aggregation level yield the minimum error compared to the other configurations. Therefore, in the rest of the experiments, we used 2 hidden units and 20 minutes of aggregation level.

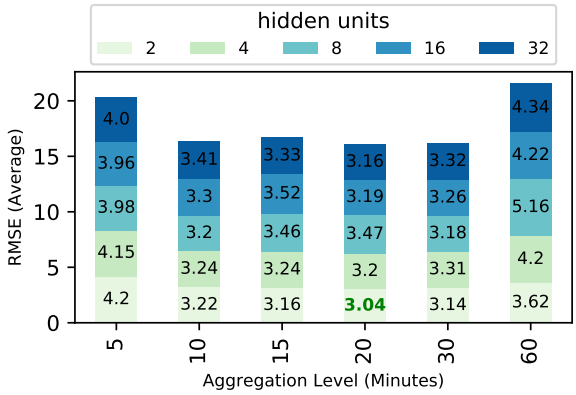


Fig. 8. Comparison of Average RMSE using different number of hidden units with various aggregation levels (d epochs).

B. Comparison of Single and Distributed Models

This experiment evaluated the proposed distributed model learning effectiveness and compared it with a single standalone learning model ($N1$). We focused on two different scenarios, where the processes can work in multiple CPUs in the same place (e.g., Tensorflow working with multiple CPUs in the same machine), or a scenario where CPUs are disaggregated and become independent of each other (e.g., different Edge nodes cooperating). The best hyper-parameter configurations

TABLE I
COMPARISON OF ERROR AND TRAINING TIME WITH FIXED NUMBER OF EPOCHS (FIXED EP) AND PROPOSED NUMBER OF EPOCHS (PROP EP).

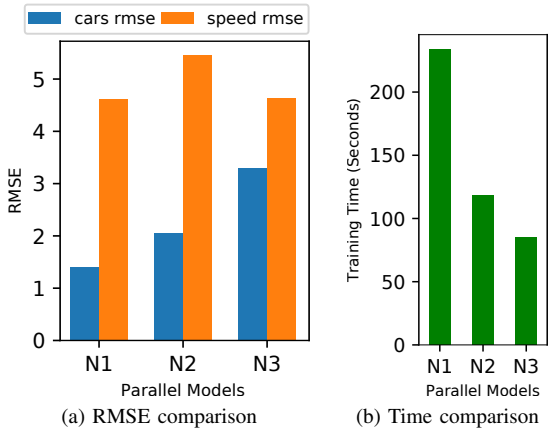
Model	RMSE (Cars)		RMSE (Speed)		Time (Sec)	
	Fixed Ep	Prop. Ep	Fixed Ep	Prop. Ep	Fixed Ep	Prop. Ep
N1	1.40	1.40	4.62	4.62	233.71	233.71
N2	3.25	2.05	7.84	5.45	237.79	118.32
N3	6.25	3.29	6.83	4.64	229.53	84.69

TABLE II
COMPARISON OF CHANGE FACTOR IN ERROR AND TRAINING TIME FOR N2 AND N3 WITH N1.

Model	+RMSE (Cars)	+RMSE (Speed)	+Time (s)
N2	0.65	0.83	-115.39
N3	1.89	0.02	-149.02

from the previous experiment were used in this experiment, with the addition that when distributing the data to be modeled, we are applying the $1/N$ factor to the number of training epochs as the "proposed epochs".

Table I shows the comparison of RMSE for the number of cars and speed with the fixed and proposed number of epochs for a single model (N1) and distributed models N2 and N3. For training distributed models N2 and N3, we used 2 and 3 Edge nodes, respectively. Whereas for N1, we used only one Edge node. We observed that for N2 and N3, using a fixed number of epochs increases RMSE due to over-fitting the model. We also observed that training time did not change even when we distribute the input data to be processed by more than one model. This occurred because the number of epochs stays the same for each configuration. However, we observed a significant decrease in training time when the number of epochs is obtained by dividing the optimal number of epochs for $N1$, by the number of parallel models. We observed that RMSE is slightly increased in estimating the number of cars, while it remains almost stable for estimating cars' speed.



(a) RMSE comparison (b) Time comparison

Fig. 9. Comparison of RMSE vs. training time for different parallel models(Nx) with number of epochs divide by x where x=1,2,3.

Figure 9 shows the accuracy for estimating the number of cars and speed for using N1, N2, and N3. Figure 9a shows that the accuracy in the estimate of the number of cars slightly decreases with the increase in the number of distributed models used to train the input data. This occurs because models are trained on fewer data and are more specific to a particular input set. This was why we observed this behavior when we combined them in the final prediction model. However, this does not affect the accuracy of predicting the speed of cars, and it somehow remains stable regardless of the number of models used to train the input data set. We also observed losing some accuracy on average, but we are saving more than 50% of training time when we used parallel models, as shown in Figure 9b.

With respect to the speed-up comparison for parallel models (N2, N3) with N1, Table II shows the improvement factor for error and time. We observed that when we distribute the input data set to be processed by 2 models. There is a decrease of 115.39 seconds in training time with an increase of 0.65 and 0.83 of RMSE for the number of cars and speed. Similarly, we observed a reduction of training time when using three parallel models and an increase of 1.89 and 0.02 of RMSE for the number of cars and speed estimations.

C. Evaluation on Low-power Architectures

In this experiment, we compared the proposed solution's effectiveness on low-power and resource constraint devices designed for the Edge, like the *Raspberry Pi* model 3B and the *NVIDIA Jetson* model Nano. Such devices are built for consuming less than 12W and embed low CPU and GPU computing resources. Raspberry Pi is used for general purposes while the Jetson integrates a GPU towards AI and neural network computing on the Edge and smart devices.

Testing the grid configurations for *Time Aggregation* vs. *Hidden Units* on the Raspberry Pi and the Jetson Nano, we observed a noticeable increase in execution time in comparison with the single-CPU Xeon. Still, the training plus validation time is below 30 minutes for nearly a week worth of data. We tested 4, 32 as 512 Batch Sizes (BS), i.e., the number of samples used for each training step in the neural network to check the Jetson GPU's possible advantages due to data bandwidth. The bigger the batch size, the more we profit from the GPU's parallelism up to a certain point. The number of epochs is fixed at 94, to compare the performance of identical training processes, and the steps (iterations) per epoch are proportional to the batch size (200 steps/epoch for BS = 4, 25 steps/epoch for BS = 32, 1 step/epoch for BS=512). The objective was to test the method's performance on low-powered devices with different properties while maintaining the error (that may vary when modifying the batch-size). As a comparison metric, we show the milliseconds per step and the seconds per epoch. When computing the average milliseconds/step, the first epoch was excluded as it carries the overhead on warm-up around ×4 the average epoch. Figure 10 shows the performance in times per step for the different configurations of the GRU in the different used technologies, for the training time with a common and proper configuration found for the GRU on the single-CPU Xeon, the Raspberry Pi ARM-based CPU, and the ARM-based and GPU enhanced Jetson.

From this experiment, we concluded that our method is fully fit for use on low-power or resource-constrained devices,

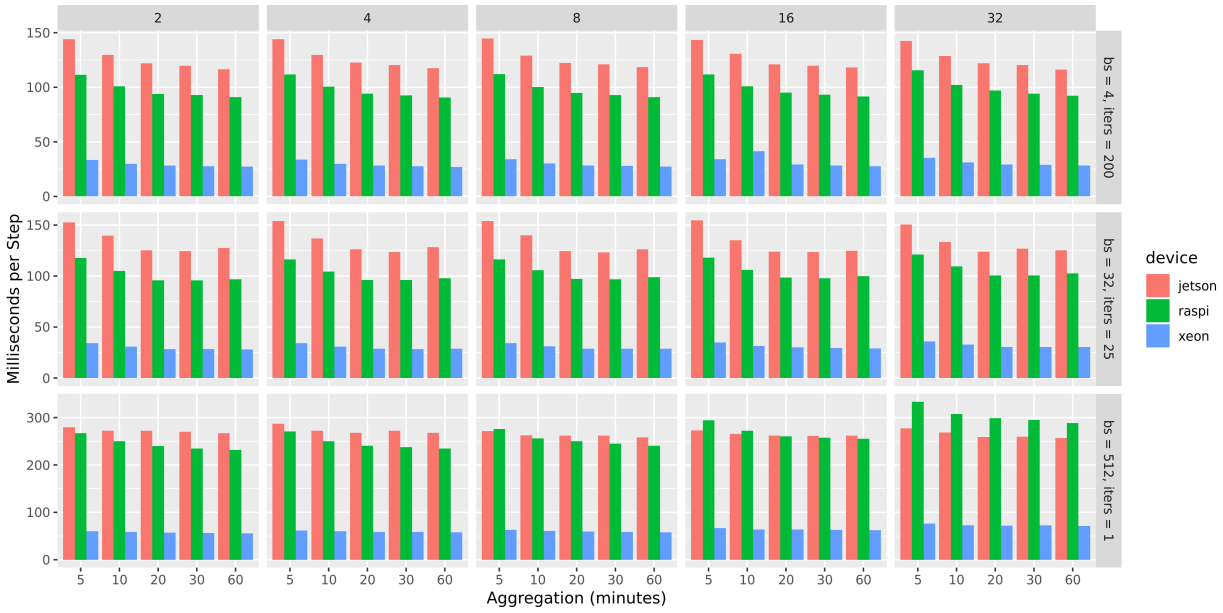


Fig. 10. Time comparison for configurations in low-power devices vs. single CPU Xeon ref., for each amount of hidden units.

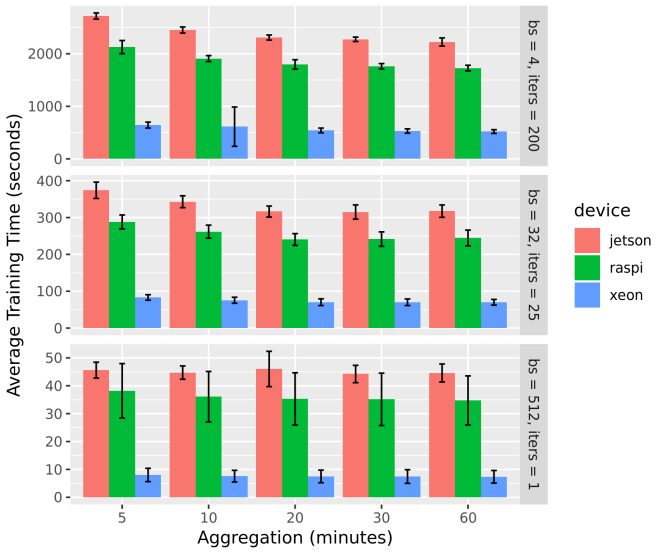


Fig. 11. Average modeling time on different Edge devices.

as training times take at maximum half an hour for a model representing around six days. Moreover, we noticed that the GPU at the Jetson Nano does not provide improvement for the kind of data until the batch size reaches larger sizes.

Figure 11 shows the absolute training time for the 3 devices, where the single Xeon outperforms the low-powered ones but for no more than a factor of 4, and how in scenarios requiring large memory bandwidth (low data aggregation and large batch sizes), the GPU starts chasing CPU execution times.

D. Comparison with Baseline Methods

To conclude the evaluations, we provide a comparison of the proposed method with previous and other simplistic models used for estimating the traffic data. We compared our solution with VAR (Vector Auto-Regression), a classic time-series analysis method, and CRBMs (Conditional Restricted Boltzmann Machines) as used in previous works [6].

TABLE III
COMPARISON WITH BASELINE MODELS VAR AND CRBMS AS N = 1 (5 MIN AGGREGATION). NOTICE THAT THE RESULTS ARE FOR THE OVERALL BEST CONFIGURATIONS FOUND.

Method	RMSE	
	Cars	**Speed**
VAR	4.99	7.44
CRBM	2.03	5.68
GRU	1.76	6.17

As we can see in Table III, the proposed solution based on GRU outperforms VAR and provided comparable performance with CRBM when the granularity is set to 5 minutes. In Figure 12 we can observe that GRU is slightly better than CRBM when granularity is finer, as seen in Figure 7. Both kind of neural network perform well in our framework. However, due to our particular interest in finer granularity, GRU is the chosen method for this work. For other experiments with different data sets, both methods should be compared in order to select the final model.

E. Discussion

Computing devices over the Edge are power and resource constrained as compared to the resource available in data centers. Building intelligent solutions requiring training compute-intensive DNN models introduced the challenge of efficiently utilizing the available Edge devices. In this work, we have addressed this challenge and proposed a system that distributes the compute-intensive ML tasks to the available Edge device while obtaining an accuracy comparable to models trained on the single machine. Our solution is capable of stopping the model training to achieve acceptable performance dynamically.

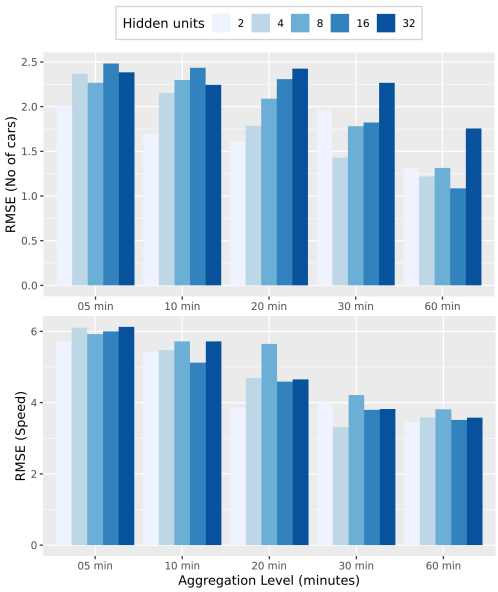


Fig. 12. Error vs. Hidden units vs. Time aggregation for number of cars and speed estimation on CRBMs.

Our experimental evaluations compared a distributed learning model with a single-model approach and other baseline solutions for traffic forecasting. The results show the potential of the proposed solution for Edge and Fog platforms.

We consider splitting the ML modeling process attractive in scenarios where we must reduce the training time without losing accuracy and constraint to avoid the task offloading to the cloud data centers. In such a situation, each device can take care of their local data and send only the trained model configuration to coordinate with other devices for building the model for generalizing the estimations. This solution is very useful in Edge computing environments in which we have low-power devices scattered. Our proposed solution will help many scenarios, including smart cities, traffic management and planning, the Internet of Things, and intelligent surveillance.

VI. Conclusions and Future Work

The presented work focused on performing predictive analytics on the Edge, using urban traffic prediction as an essential use case scenario relevant to Smart Cities applications. Given the amount of data generated on the Edge, not only in volume but also in time, moving modeling and analytics near the data may be a good compromise in front of Cloud models, where data must be massively pushed northbound. Of course, it must be understood that Deep Learning and other analytics processes are usually designed for high-performance computing environments. In contrast, the Edge front is commonly composed of low-power devices with scarce computing resources.

In this paper, we proposed an experimentally-evaluated method based on FL to move data analytics processing to Edge. The learning tasks are distributed to multiple Edge nodes, with limited computing resources, in a manner that each node processes its own local data. In so doing, we attempted to balance training time and model accuracy as a function of data distribution. Experiments showed that the tested data-sets, provided by a road-service car fleet from Barcelona, can be learned with acceptable accuracy although being unstable on different previous tested techniques. Also, for each configuration of NNs, there exists a multi-dimensional trade-off between the time spent on training, the distribution of data and parallelization of the model training process, and the previous aggregation of collected data to be trained, creating an interesting problem on how good we can model traffic against how much available time/resources are on given Edge scenarios.

The presented solution highlights future research and innovation opportunities on Smart City applications, capable of providing services near-data and near-users without abusing network hierarchies and Cloud resources. While this work focused on a specific type of NNs, other statistical and ML can be applied, more suitable for particular scenarios far from urban traffic. Also, more complex architectures for distributing machine learning processes and automation of autonomous learning can be applied, focusing on better decisions when having time/resources for smart management of the device and the data pipeline. Another interesting aspect to cover is the use of MCUs with FL. With this, ultra-low powered devices would be able to collaborate in training and provide models with a wider knowledge on data from their neighboring MCUs.

Acknowledgments

This work has been partially supported by the Spanish Government (contract PID2019-107255GB), the Generalitat de Catalunya (contract 2014-SGR-1051), the University of Padova, and Severo Ochoa CoE (SEV-2015-0493-16-5). The statements made herein are solely the responsibility of the authors.

References

[1] M. Abdullah, W. Iqbal, A. Mahmood, F. Bukhari, and A. Erradi, "Predictive autoscaling of microservices hosted in fog microdata center," *IEEE Systems Journal*, 2020.

[2] S. Ali, A. Ashraf, S. B. Qaisar, M. Kamran Afridi, H. Saeed, S. Rashid, E. A. Felemban, and A. A. Sheikh, "Simplimote: A wireless sensor network monitoring platform for oil and gas pipelines," *IEEE Systems Journal*, vol. 12, no. 1, pp. 778–789, 2018.

[3] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: A survey," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4158–4168, 2019.

[4] G. Premsankar, M. Di Francesco, and T. Taleb, "Edge computing for the internet of things: A case study," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1275–1284, 2018.

[5] G. Plastiras, M. Terzi, C. Kyrkou, and T. Theocharidcs, "Edge intelligence: Challenges and opportunities of near-sensor machine learning applications," in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 2018, pp. 1–7.

[6] J. L. Pérez, A. Gutierrez-Torre, J. L. Berral, and D. Carrera, "A resilient and distributed near real-time traffic forecasting application for fog computing environments," *Future Generation Computer Systems*, vol. 87, pp. 198 – 212, 2018.

[7] C. Savaglio and G. Fortino, "A simulation-driven methodology for iot data mining based on edge computing," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 2, pp. 1–22, 2021.

[8] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *ArXiv*, vol. abs/1602.05629, 2016.

[9] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[10] S. Dey, A. Chakraborty, S. Naskar, and P. Misra, "Smart city surveillance: Leveraging benefits of cloud data stores," in *37th Annual IEEE Conference on Local Computer Networks-Workshops*. IEEE, 2012, pp. 868–876.

[11] K. Su, J. Li, and H. Fu, "Smart city and the applications," in *2011 international conference on electronics, communications and control (ICECC)*. IEEE, 2011, pp. 1028–1031.

[12] M. Castro, A. J. Jara, and A. F. Skarmeta, "Smart lighting solutions for smart cities," in *2013 27th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2013, pp. 1374–1379.

[13] S. Yi, Z. Hao, Q. Zhang, Q. Zhang, W. Shi, and Q. Li, "Lavea: Latency-aware video analytics on edge computing platform," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*. ACM, 2017.

[14] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[15] A. Sergeev and M. Del Balso, "Horovod: fast and easy distributed deep learning in tensorflow," *arXiv preprint arXiv:1802.05799*, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *arXiv preprint arXiv:1202.2745*, 2012.

[18] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[19] B. Hu, Y. Gao, L. Liu, and H. Ma, "Federated region-learning: An edge computing based framework for urban environment sensing," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–7.

[20] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," 2018.

[21] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223–1231.

[22] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, "Building high-level features using large scale unsupervised learning," *arXiv preprint arXiv:1112.6209*, 2011.

[23] N. Strom, "Scalable distributed dnn training using commodity gpu cloud computing," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 1488–1492.

[24] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, and N. Andrew, "Deep learning with cots hpc systems," in *International conference on machine learning*, 2013, pp. 1337–1345.

[25] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint:1706.02677*, 2017.

[26] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Opportunistic spatio-temporal event processing for mobile situation awareness," in *Proceedings of the 7th ACM international conference on Distributed event-based systems*. ACM, 2013, pp. 195–206.

[27] L. Yu, Z. Li, J. Liu, and R. Zhou, "Resources sharing in 5g networks: Learning-enabled incentives and coalitional games," *IEEE Systems Journal*, pp. 1–12, 2019.

[28] A. Marchisio, M. A. Hanif, F. Khalid, G. Plastiras, C. Kyrkou, T. Theocharides, and M. Shafique, "Deep learning for edge computing: Current trends, cross-layer optimizations, and open research challenges," in *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2019, pp. 553–559.

[29] B. Sudharsan, J. G. Breslin, and M. I. Ali, "Rce-nn: a five-stage pipeline to execute neural networks (cnns) on resource-constrained iot edge devices," in *Proceedings of the 10th International Conference on the Internet of Things*, 2020, pp. 1–8.

[30] R. Sanchez-Iborra and A. F. Skarmeta, "Tinyml-enabled frugal smart objects: Challenges and opportunities," *IEEE Circuits and Systems Magazine*, vol. 20, no. 3, pp. 4–18, 2020.

[31] S. Lee and S. Nirjon, "Neuro.zero: A zero-energy neural network accelerator for embedded sensing and inference systems," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, ser. SenSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 138–152. [Online]. Available: https://doi-org.recursos.biblioteca.upc.edu/10.1145/3356250.3360030

[32] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[33] M. Naphade, D. C. Anastasiu, A. Sharma, V. Jagrlamudi, H. Jeon, K. Liu, M.-C. Chang, S. Lyu, and Z. Gao, "The nvidia ai city challenge," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017, pp. 1–6.

[34] "Nvidia autonomous machines: Jetson nano," July 2019. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines

[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[36] F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.