# Communication-Efficient Federated Learning for Connected Vehicles with Constrained Resources

Shuaiqi Shen[1], Chong Yu[1], Kuan Zhang[1], Xi Chen[2], Huimin Chen[2], Song Ci[3]
1.Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, NE, USA, 68182
2. Beijing Electric Power Research Institute, State Grid Corporation of China
3. State Key Laboratory for Safety Control and Simulation of Power Systems and Large Power Generation Equipment, Tsinghua University, Beijing, P. R. China, 100084
Email: sshen@huskers.unl.edu, cyu6@huskers,unl.edu, kuan.zhang@unl.edu, xchenaz@163.com, hmchen@sina.com, sci@tsinghua.edu.cn

2021 International Wireless Communications and Mobile Computing (IWCMC) | 978-1-7281-8616-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IWCMC51323.2021.9498677

***Abstract*—With the upcoming next generation wireless network, vehicles are expected to be empowered by artificial intelligence (AI). By connecting vehicles and cloud server via wireless communication, federated learning (FL) allows vehicles to collaboratively train deep learning models to support intelligent services, such as autonomous driving. However, the large number of vehicles and increasing size of model parameters bring challenges to FL-empowered connected vehicles. Since communication bandwidth is insufficient to upload full-precision local models from numerous vehicles, model compression is usually conducted to reduce transmitted data size. Nevertheless, conventional model compression methods may not be practical for resource-constrained vehicles due to the increasing computational overhead for FL training. The overhead for downloading global model can also be omitted by existing methods since they are originally designed for centralized learning instead of FL. In this paper, we propose a ternary quantization based model compression method on communication-efficient FL for resource-constrained connected vehicles. Specifically, we firstly propose a ternary quantization based local model training algorithm that optimizes quantization factors and parameters simultaneously. Then, we design a communication-efficient FL approach that reduces overhead for both upstream and downstream communications. Finally, simulation results validate that the proposed method demands the lowest communication and computational overheads for FL training, while maintaining desired model accuracy compared to existing model compression methods.**

***Index Terms*—Federated learning, model compression, connected vehicles, ternary quantization**

## I. Introduction

Benefited from the evolution of Internet of Things (IoT) and the ubiquity of Artificial Intelligence (AI), the techniques for autonomous driving are emerging in the next generation of wireless network. Connected via wireless communication, vehicles can cooperate with other nearby vehicles and the edge of network to improve driving safety and traffic efficiency [1]. As shown in Fig. 1, connected vehicles primarily support four types of intelligent services, including environment perception, map building, path planning and motion control [2]. These services are driven by AI to generate deep learning models for real-time decision making based on the environment and traffic-related data collected by connected vehicles. To leverage the sheer volume of data distributed among a large number of vehicles in the network, federated learning (FL) is utilized to allow various vehicles to collaboratively train a deep learning model without sharing their private data. Each vehicle maintains its own local model and updates the model parameters based on local training data with certain optimizer, such as stochastic gradient descent (SGD) [3]. Then, the local model parameters are uploaded to the cloud server and aggregated to generate a global model, whose parameters are downloaded by vehicles to improve their local models. This process iterates until the training outcome converges. Empowered by FL, connected vehicles can extract knowledge about autonomous driving from distributed data, while preserving the privacy of users' driving behaviors.

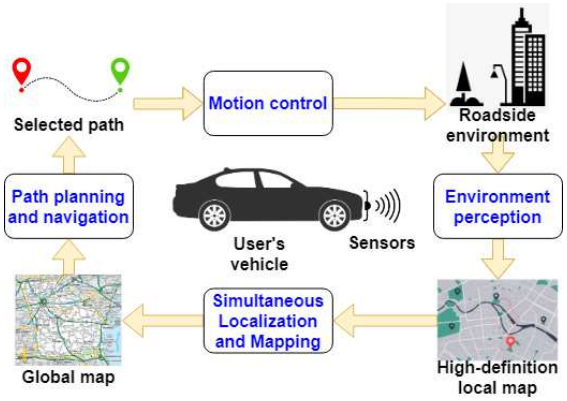



Fig. 1. Intelligent services for connected vehicles to realize autonomous driving via wireless communication

However, the massive communication and computational overheads are still major issues for FL-related applications. Compared with other distributed learning methods, FL-empowered connected vehicles encounter several unique challenges. Firstly, communication overhead becomes a major bottleneck for FL due to the rapidly increasing model complexity and the number of connected vehicles. Gigabytes of data containing millions of parameters need to be transmitted to complete a full model update in each training iteration [4]. Since communication bandwidth is usually limited for connected vehicles, FL may become impractical for services that require timely responses from deep learning models.

Model compression is necessary to reduce the communication overhead for transmitting model parameters in FL. Secondly, the computational resources are constrained for connected vehicles to deploy common model compression methods [5]. Due to the limited computing power and battery capacity of most onboard platforms [6], the local training procedure should also be simplified to alleviate the addition overhead caused by conducting model compression. Thirdly, both upstream and downstream communication overheads should be considered between connected vehicles and cloud server. A complete iteration of FL contains uploading and downloading of model updates. As conventional model compression methods [7] are designed for centralized learning, they only reduce the data size of local models uploaded from client to server.

To achieve communication-efficient FL for connected vehicles, model compression methods are integrated with FL to reduce the amount of parameters transmitted in training iterations. Existing model compression methods, such as neural network pruning [8], low-rank factorization [9], convolutional filters [10] and knowledge distillation [11], simplify the deep learning models after local training is completed. These methods are difficult to be deployed on connected vehicles with constrained resources, since the additional overhead is not alleviated for compressing local model parameters. A multi-objective optimization method is proposed to simultaneously maximize the model accuracy while minimizing model complexity [12]. Communication efficiency of FL can also be improved by uploading only parameters that lead to essential global model update [13]. Nevertheless, these methods only compress local models uploaded from vehicles to cloud server, such that the overhead reduction of downstream communication is omitted. Therefore, a compression method is necessary for connected vehicles to address the communication and computational overheads caused by FL.

In this paper, we propose a model compression method for FL-empowered connected vehicles with constrained resources. The proposed method reduces the complexity of local model training while improving the efficiency of both upstream and downstream communications simultaneously. Specifically, the contributions of this paper are three-fold.

(1) We propose a ternary quantization based algorithm for local model training to reduce the number of model parameter values. The proposed algorithm is integrated with local model training, so that the quantized values of parameters are optimized for different local models. Since the model parameters are quantized before being updated with SGD, less computational overheads are demanded to compute loss functions and parameter gradients during local training. This alleviates the addition overhead caused by the proposed method and makes FL more adaptive to the resource-constrained vehicles with limited computing power and battery capacity.

(2) We propose a ternary quantization based FL approach to reduce the overheads of upstream and downstream communications. The deep learning model complexity can be reduced for both local models updated in connected vehicles and global model aggregated on cloud server. In addition to communication efficiency, the quantized global model parameters also increase the difficulty to deduce local model parameters from vehicles, so that privacy preservation is improved.

(3) We conduct extensive simulation to evaluate the proposed method compared with existing model compression methods by training multiple popular deep learning models. The models are trained based on battery data collected from electric vehicles (EVs) [14] to simulate FL-empowered battery status estimation in connected vehicles. The simulation results validate that our proposed method achieves the lowest communication and computational overhead while maintaining desired model accuracy.

The remainder of the paper is organized as follows. Section II reviews related works of model compression methods for FL. In Section III, we propose the ternary quantization based algorithm and collaborative FL approach. Simulation results are presented in Section IV. Finally, conclusions are drawn and future directions are identified in Section V.

## II. Related Works

To improve the communication efficiency of FL, existing methods can be categorized into three types: local updating, sparsification and quantization. Firstly, local updating methods aim to reduce the frequency of communication between the clients (i.e., connected vehicles) and server. Instead of uploading and downloading model parameters in every training iteration, the clients can perform multiple local updates before uploading their parameters for aggregation [15]. This tends to have slight influence on convergence rate of FL as long as the data is independently and identically distributed among clients. The number of total communication rounds for FL training can also be reduced by decomposing the global objective into subproblems that can be solved in parallel in each iteration [16]. Secondly, sparsification methods restrict the updated parameters of local model training to a small subset to obtain sparse information communicated from clients to server. For example, for SGD-based FL, only gradients whose magnitudes exceed a predetermined threshold can be uploaded to server, while the others with less magnitudes accumulate as residual [7]. The amount of uploaded parameters can also be determined by a fixed sparsity factor representing the portion of parameters selected from the full set [17]. Finally, quantization methods alleviate communication overhead of FL by restricting updated parameters to a reduced set of values. Instead of decreasing the number of communicated parameters, quantization methods reduce the information entropy carried in each communication round. Existing quantization strategies include binary sign [18], universal vector quantization [19], stochastic gradient decomposition [20] and so on.

However, the aforementioned methods for communication-efficient FL have limitations to be deployed on resource-constrained connected vehicles. Local updating methods leverage distributed data processing to reduce communication overheads among vehicles and server, but allocate heavier burden on local model training. Since onboard platforms of connected vehicles have limited computing power, massive

local updating may instead extend the communication delay of each iteration. In addition, sparsification methods only update a small subset of parameters depending on their gradient magnitudes, such that the model compression rate is quite sensitive to the predetermined threshold or sparsity factor. Since the deep learning models utilized in various intelligent services of connected vehicles are highly diverse, adjusting the thresholds for sparsification methods can be quite difficult and inflexible. Furthermore, existing quantization methods only focus on the compression of local models, while the effective quantization on global model parameters after aggregation is in lack. The overhead of downstream communication can hardly be alleviated with conventional model compression methods. A novel parameter quantization method is necessary for FL-empowered connected vehicles. Reducing communication overheads for both uploading and downloading becomes critical, while integrating model compression with local model update has a potential to improve training efficiency.

## III. Ternary Quantization Based Model Compression for Federated Learning

In this section, we propose an adaptive ternary quantization algorithm that optimizes the quantized values for local model parameters of different vehicles. Then, a communication-efficient FL approach is provided that reduces the overheads of both upstream and downstream parameter transmission.

### A. Adaptive Ternary Quantization with Gradient Descent

The communication efficiency of FL is determined by the size of transmitted data between connected vehicles and cloud server. By quantizing the model parameters from continuous values to discrete values, fewer bits are needed to represent the deep learning model uploaded or downloaded during FL. Since the scales and sparsity of model parameters may greatly differ from vehicle to vehicle, normalization is necessary for parameters to avoid bias towards larger magnitudes during quantization. Suppose a deep learning model trained by FL has $L$ layers, and each layer has $d_l$ dimensions where $l = 1, \ldots, L$. Parameter normalization is firstly conducted layer by layer as $\theta_l^n = g(\theta_l)$, where $\theta_l$ is the parameters of layer $l$ and $g$ is a function that normalizes a vector into the range of $[-1, 1]$. Then, ternary quantization assigns the normalized parameters to be positive, zero, or negative layer by layer

$$\theta_l^t = \begin{cases} \omega_l, & \theta_l^n > \Delta_l \\ 0, & |\theta_l^n| \leq \Delta_l \\ -\omega_l, & \theta_l^n < -\Delta_l \end{cases} \tag{1}$$

where $\theta_l^t$ is the quantized parameter of layer $l$, $\omega_l$ is a quantization factor and $\Delta_l$ is the threshold for quantization. Both quantization factors $\omega_l$ and $\Delta_l$ should be positive, which determine the scale and sparsity of model parameters after quantization respectively, so that both factors have great impacts on the accuracy of compressed models. The quantization threshold is computed based on parameters in the same layer

$$\Delta_l = \frac{T_l}{d_l \times d_{l+1}} \sum_{i=1}^{d_l} \sum_{j=1}^{d_{l+1}} |\theta_{lij}^n|, \tag{2}$$

where $\theta_{lij}^n$ is the entry of $i^{th}$ row and $j^{th}$ column of the parameter matrix $\theta_l^n$ for the full connection between layer $l$ and $l+1$. The factor $T_l$ is another quantization factor that determines the upper bound of the quantization threshold

$$\begin{aligned} \Delta_l =& T_l \times \frac{1}{d_l d_{l+1}} \sum_{i=1}^{d_l} \sum_{j=1}^{d_{l+1}} |\theta_{lij}^n| \\ \leq& T_l \times \frac{1}{d_l d_{l+1}} (d_l d_{l+1} \times \max|\theta_{lij}^n|) \\ \leq& T_l \times \max|\theta_{lij}^n| \leq T_l, \end{aligned} \tag{3}$$

such that the boundary of parameter values for positive, zero and negative ranges can be adaptive for different layers. Due to the divergence among local models generated by different connected vehicles, the two factors $\omega_l$ and $T_l$ should be adaptive to the parameter distribution to preserve the highest model accuracy after quantization.

To obtain the optimal quantization factors for different vehicles and different layers in local models, we propose a gradient descent based algorithm to update factor values iteratively. Since most existing FL methods [21]–[23] also apply gradient descent to optimize local model parameters, the optimization of quantization can be integrated with local training easily in each round of FL. Instead of conducting additional computation for model compression after local model being generated, the proposed method simplifies local training procedure by quantizing parameters before optimizing them with SGD. With previously quantized parameters, computing loss functions and gradients becomes more efficient than using the continuous parameter values. This reduces the computational overhead of resource-constrained vehicles, so that their limited computing power and battery capacity are no longer bottlenecks to realize communication-efficient FL. We denote the loss function of local model training for vehicle $k$ as

$$J_k(\theta^t) = \frac{1}{|D_k|} \sum_{i=1}^{|D_k|} f(x_i, y_i; \theta^t), \tag{4}$$

where $D_k$ is the local training dataset of vehicle $k$ containing pairs of features and objective value $(x_i, y_i) \in D_k, i = 1, \ldots, |D_k|$. Suppose totally $N$ connected vehicles are involved in FL, the global loss function can be calculated

$$J(\theta^t) = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} J_k(\theta^t). \tag{5}$$

Based on the loss function value, the gradients of quantization factor $\omega_l$ for vehicle $k$ can be calculated

$$\frac{\partial J_k}{\partial \omega_l} = \sum_{(i,j) \in \{(i,j) | \theta_{lij}^n > \Delta_l\}} \frac{\partial J_k}{\partial \theta_{lij}^t}. \tag{6}$$

The gradient of factor $T_l$ is more difficult to obtain, since $T_l$ determines the quantization threshold instead of parameter values. However, $T_l$ has a strong correlation with $\omega_l$, as they determine the distribution of quantized model parameters together. When $\omega_l$ increases, the magnitudes of quantized

parameters becomes larger in the corresponding layer, such that $T_l$ should also increase to assign fewer parameters as $\omega_l$ or $-\omega_l$. This guarantee that the average value of quantized parameters remains stable for different iterations of model training. To this end, we assign the update of $T_l$ in the same direction as the gradient of $\omega_l$. After quantization, the model parameters are updated with gradients to the local loss function for vehicle $k$

$$\frac{\partial J_k}{\partial \theta_l} = \frac{\partial J_k}{\partial \theta_l^t} \times \frac{\partial \theta_l^t}{\partial \theta_l} = \begin{cases} 1 \times \frac{\partial J_k}{\partial \theta_l^t}, & |\theta_l^n| \leq \Delta_l \\ \omega_l \times \frac{\partial J_k}{\partial \theta_l^t}, & otherwise. \end{cases} \quad (7)$$

**Algorithm 1:** Gradient Descent based Ternary Quantization Algorithm (GDTQ) for FL

**Input:** $\Theta = \{\theta_1, \ldots, \theta_L\}$, $where\ \theta_l \in R^{d_l \times d_{l+1}}$; $\omega_l, T_l$ for all layers $l = 1, \ldots, L$; local dataset $(x_i, y_i) \in D_k$ for client $k$; local training iterations $I_{max}$;
**Output:** quantized parameters $\Theta^t = \{\theta_1^t, \ldots, \theta_L^t\}$; updated quantization factors $\omega_l, T_l$;
**Initialization:** learning rate $\alpha_1, \alpha_2, \alpha_3$, loss function $f$, normalization function $g$, $\Theta^t = \Theta$;
**for** $iter = 1$ *to* $I_{max}$ **do**
  $J_k(\Theta^t) = \frac{1}{|D_k|} \sum_{i=1}^{|D_k|} f(x_i, y_i; \Theta^t)$;
  **for** $l = 1$ *to* $L$ **do**
    $\theta_l^n = g(\theta_l)$;
    $\Delta_l = \frac{T_l}{d_l \times d_{l+1}} \sum_{i=1}^{d_l} \sum_{j=1}^{d_{l+1}} |\theta_{lij}^n|$;
    quantize $\theta_l^n$ with Eqn. 1 to obtain $\theta_l^t$;
    $\omega_l = \omega_l + \alpha_1 \frac{\partial J_k}{\partial \omega_l}$;
    $T_l = T_l + \alpha_2 sign(\frac{\partial J_k}{\partial \omega_l})$;
    $\theta_l^n = \theta_l^n + \alpha_3 \frac{\partial J_k}{\partial \theta_l}$;
  **end**
**end**

The proposed ternary quantization algorithm is summarized in Algo. 1, where the quantization rules are adaptive to parameter distributions of different vehicles and neural network layers. The model compression procedure is integrated with local model training to reduce the upstream communication overheads for resource-constrained vehicles. After a certain number of local training iterations, the quantized model parameters and quantization factors are uploaded to cloud server for aggregation. However, the aggregated parameters of global model can have more diverse values due to the various quantization scales applied in different local models. Larger data size is needed to carry the increased information entropy of global model for downstream communication than local models. To address this issue, the next subsection discusses the proposed overall FL approach that achieves efficiency for both upstream and downstream data transmission.

### *B. Communication-efficient Federated Learning*

The proposed FL approach is illustrated in Fig. 2. Each vehicle normalizes and quantizes its local model parameters during the local training as summarized in Algo. 1. After

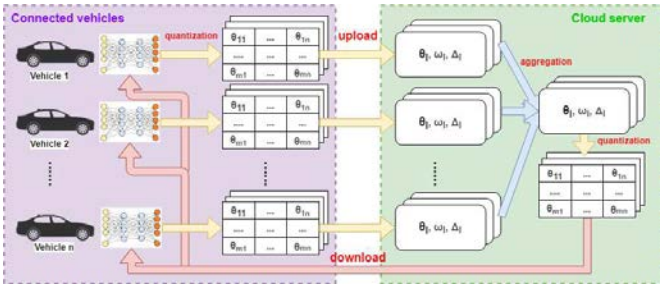



Fig. 2. Proposed communication-efficient FL for connected vehicles

**Algorithm 2:** Ternary quantization based FL

**Initialization:** Global model parameters $\Theta_0$, maximum number of FL rounds $R_{max}$, number of involved vehicles $N$, local training iteration $I_{max}$;
**for** $r = 1$ *to* $R_{max}$ **do**
  **Vehicles do in parallel:**
  **for** $k = 1$ *to* $N$ **do**
    load dataset $D_k$;
    $\Theta_r = \Theta_{r-1}$;
    **for** $l = 1$ *to* $L$ **do**
      $T_l = 0.7$;
      $\omega_l = \frac{1}{|I_p|+|I_n|} \sum_{(i,j) \in I_p \cup I_n} |\theta_{lij}|$;
    **end**
    $\Theta_r^t = GDTQ(\Theta_r, T, \omega, D_k, I_{max})$;
    upload $\Theta_r^t$ to server with updated factors $T, \omega$;
  **end**
  **Server does:**
  $\Theta_r = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} \Theta_{k,r}^t$;
  **for** $l = 1$ *to* $L$ **do**
    $\omega_l = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} \omega_{l,k}$;
    $T_l = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} T_{l,k}$;
    $\Delta_l = \frac{T_l}{d_l \times d_{l+1}} \sum_{i=1}^{d_l} \sum_{j=1}^{d_{l+1}} |\theta_{lij}^n|$;
    quantize $\Theta_r$ with Eqn. 1;
    broadcast quantized $\Theta_r$ to all vehicles;
  **end**
**end**

certain iterations, the quantized parameters and corresponding quantization factors are uploaded to the cloud server. Then, the global model is generated by aggregating local models as weighted average

$$\Theta = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} \Theta_k^t, \quad (8)$$

where $N$ is the total number of vehicles involved in FL, $|D_k|$ is the size of local dataset from vehicle $k$, and $\theta_k^t$ is the quantized local model parameters. New quantization factors are obtained by aggregating uploaded factors from vehicles in a similar way

$$\omega_l = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} \omega_{l,k}; \quad T_l = \sum_{k=1}^{N} \frac{|D_k|}{\sum_{k=1}^{N} |D_k|} T_{l,k}. \quad (9)$$

Finally, global model parameters are quantized with new quantization factors to be downloaded by all vehicles. The new quantization factors obtained on server are not downloaded by vehicles, since the factors need to be initialized for each round of FL to remain adaptive to the updated model parameters. At the beginning of local training, we initialize quantization threshold factor $T_l = 0.7$ for all layers $l = 1, \ldots, L$, as the range of normalized parameters is between -1 and 1. The weight factor is initialized based on updated parameters

$$\omega_l = \frac{1}{|I_p| + |I_n|} \sum_{(i,j) \in I_p \cup I_n} |\theta_{lij}|, \tag{10}$$

where $I_p = \{(i,j)|\theta_{lij} > \Delta_l\}$ and $I_n = \{(i,j)|\theta_{lij} < -\Delta_l\}$. Based on the proposed Gradient Descent based Ternary Quantization Algorithm (GDTQ), we propose the overall FL approach as summarized in Algo. 2. By quantizing local and global models, the proposed method can substantially reduce the communication overheads for both upstream and downstream data transmission. This brings an essential advantage for deploying FL in resource-constrained connected vehicles.

## IV. Performance Evaluation

In this section, we conduct extensive simulation to evaluate the proposed method based on battery operating data of EVs [14] to conduct battery status estimation. To establish an accurate estimation model for the status of onboard battery, large volume of training data need to be collected from different vehicles. Empowered by FL, connected vehicles can train the battery model collaboratively without sharing their private operating data. This is a suitable scenario to validate our proposed method on supporting communication-efficient FL for intelligent services of connected vehicles. The training dataset contains 16 features and over 130000 samples, where the objective values are classified into 10 different status for users to monitor the battery operating conditions. As shown in Fig. 3, totally 10 EVs are connected to conduct FL. The simulation is repeated for 10 echoes to obtain the average model accuracy after FL training. Two other model compression methods based on sparse ternary compression (**STC**) [5] and model pruning (**MP**) [8] are utilized in FL to be compared with the proposed method. As the performance benchmark, FL is also conducted without any model compression to show the impact of quantization on model accuracy. We train both artificial neural network (ANN) and convolutional neural network (CNN) in the simulation to validate the generality of the proposed method on different deep learning models. The simulation settings are summarized in Table. 1.

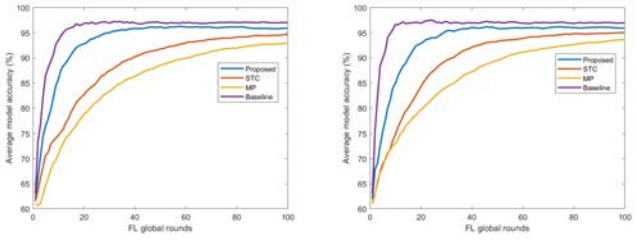
To evaluate the obtained accuracy and efficiency, average model accuracy, transmitted data size and computational time are measured throughout the FL training procedure for different deep learning models and model compression methods. As shown in **Fig. 4**, the proposed method achieves higher model accuracy than other two model compression methods. This is because our quantization factors are optimized during local training to fit the parameter updates, while the compared methods simplify local models only after training is completed.

Fig. 3. One of the EVs involved in FL for onboard battery status estimation

TABLE I
SETTINGS OF MODEL COMPRESSION ASSISTED FL

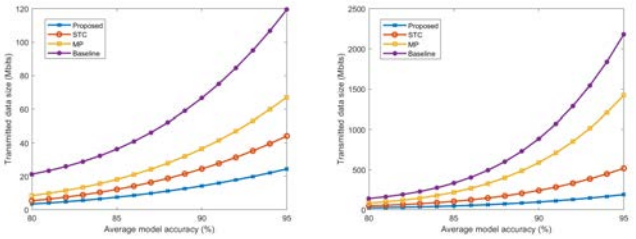
| **Simulation settings** | **Applied values** |
|---|---|
| Deep learning model | ANN, CNN |
| Model compression method | Proposed, STC, MP, no compression |
| Local training iterations | 20 |
| FL rounds | 100 |
| Clients (connected vehicles) | 10 |
| Simulation repetitions | 10 |

Although ternary quantization decreases parameter precision, the proposed method can still obtain desired accuracy close to the baseline after sufficient FL training rounds. **Fig. 5** shows the transmitted data size of model uploading and downloading. The proposed method demands the least communication overhead since it also alleviates downstream overhead, while conventional methods can only compress uploaded models when utilized in FL. **Fig. 6** shows that two compared methods cause additional computational overheads to FL to conduct model compression, while our proposed method reduces the training overhead compared to baseline. The reason is that during local training, the model parameters are firstly quantized and then updated with SGD, so that the number of operations to compute loss functions and gradients can be substantially reduced. Therefore, simulation results validate that the proposed method achieves the lowest communication and computational overheads for FL-empowered connected vehicles, while maintaining desired model accuracy close to the original FL approach without model compression.



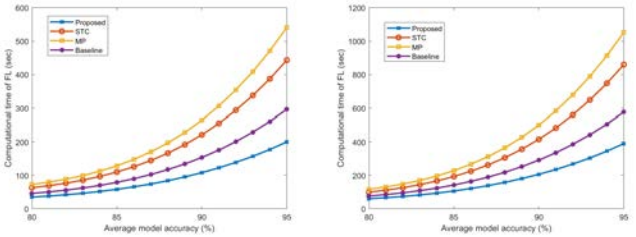

(a) (b)

Fig. 4. Model accuracy versus global rounds of FL by training: (a) ANN model; (b) CNN model




(a) (b)

Fig. 5. Communication overheads of parameter uploading and downloading versus desired model accuracy by training: (a) ANN model; (b) CNN model




(a) (b)

Fig. 6. Overall computational time of FL versus desired model accuracy by training: (a) ANN model; (b) CNN model

## V. Conclusion

In this paper, we have proposed a ternary quantization based model compression method to realize communication-efficient FL for resource-constrained connected vehicles. Specifically, we have proposed a local model training algorithm that integrates quantization with SGD based model training to reduce computational overhead brought by model compression. Then, we have designed a ternary quantization based FL approach for efficient upstream and downstream communications, so that FL can be practical for resource-constrained vehicles. Finally, simulation results validate that the proposed method can achieve the lowest communication and computational overheads while maintaining desired model accuracy. In future works, we will focus on integrating the proposed method with the methods that reduce transmission overhead from other directions, such as feature selection and client sampling, to collaboratively improve the communication efficiency for FL.

## Acknowledgment

This work is partly supported by the Science Program of State Grid Corporation of China under Grant No. 52022319005Q.

## References

[1] Y. Chen, C. Lu, and W. Chu, "A cooperative driving strategy based on velocity prediction for connected vehicles with robust path-following control," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3822–3832, 2020.

[2] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1826–1848, 2019.

[3] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[4] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[5] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[6] S. Shen, S. Ci, K. Zhang, and X. Liang, "Lifecycle prediction of second use electric vehicle batteries based on arima model," in *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2019, pp. 1–6.

[7] H. Sun, X. Ma, and R. Q. Hu, "Adaptive federated learning with gradient compression in uplink NOMA," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 325–16 329, 2020.

[8] P. Singh, V. K. Verma, P. Rai, and V. P. Namboodiri, "Acceleration of deep convolutional neural networks using adaptive filter pruning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 838–847, 2020.

[9] S. Jiang, Z. Ding, and Y. Fu, "Heterogeneous recommendation via deep low-rank sparse collective factorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 5, pp. 1097–1111, 2019.

[10] T. Li, B. Wu, Y. Yang, Y. Fan, Y. Zhang, and W. Liu, "Compressing convolutional neural networks via factorized convolutional filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3977–3986.

[11] T. Guo, C. Xu, S. He, B. Shi, C. Xu, and D. Tao, "Robust student network learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2455–2468, 2019.

[12] H. Zhu and Y. Jin, "Multi-objective evolutionary federated learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 4, pp. 1310–1322, 2019.

[13] L. Wang, W. Wang, and B. Li, "CMFL: Mitigating communication overhead for federated learning," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 954–964.

[14] S. Shen, B. Liu, K. Zhang, and S. Ci, "Toward fast and accurate soh prediction for lithium-ion batteries," *IEEE Transactions on Energy Conversion*, to appear.

[15] 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," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[16] V. Smith, S. Forte, M. Chenxin, M. Takáč, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *Journal of Machine Learning Research*, vol. 18, p. 230, 2018.

[17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[18] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 560–569.

[19] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 500–514, 2020.

[20] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2128–2142, 2020.

[21] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 8, pp. 1754–1766, 2020.

[22] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[23] H. Guo, A. Liu, and V. K. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 197–210, 2020.