Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 3, 2025

Entropy-based Adaptive Gradient Quantization in Federated Learning for Internet of Vehicles

Zhaocheng Luo

College of Information Engineering, Henan University of Science and Technology, Luoyang, 471000, Henan, China

Abstract: Federated learning for internet of vehicles builds an intelligent transportation system with real-time responsiveness and intelligent collaborative training of high-quality models by integrating traffic data between vehicle nodes, roadside units, and infrastructure. As the internet of vehicles architecture continues to expand, frequent gradient data interactions between roadside units and vehicle nodes lead to increased uplink channel load and communication delay in federated learning systems. To alleviate the communication delay problem, existing works propose gradient quantization algorithms to reduce the communication bandwidth overhead by reducing the transmission of redundant data. However, the existing gradient quantization algorithms' undifferentiated discarding of gradient data leads to a reduction in the accuracy of the aggregation model. To balance model accuracy and communication overhead, we propose an entropy-based adaptive gradient quantization for federated learning (eaqfed). The eaqfed dynamically adjusts the gradient quantization level through the entropy property during model updating to maintain model accuracy while reducing communication cost.

Keywords: Internet of Vehicles, Federated Learning, Gradient Quantization.

1. Introduction

Internet of vehicles (*iov*) integrates traffic data from multiple sources, optimizes traffic flow scheduling, improves driving safety, and gradually develops into an important component of intelligent transportation systems [1]. However, since the data collected by vehicles often involves users' private information, users are usually reluctant to upload the data to untrusted servers to participate in centralized training, and this concern triggers the data silo problem. To address the problem, federated learning (*fl*) is introduced in *iov* by performing model training locally and uploading only the gradient information of model updates, preventing raw data from leaving the local area and effectively protecting user privacy [2]. In the *fl* for *iov*, vehicles are usually equipped with lightweight convolutional neural networks, which can perform tasks such as target recognition and target detection, and realizing the application of technologies such as driverless driving, thus gradually evolving into intelligent terminal nodes (hereafter referred to as vehicle nodes). However, as the scale of *iov* expands, the amount of data grows exponentially, and the frequent data interactions between vehicle nodes and roadside units and other devices lead to increased uplink channel loading, which in turn brings about communication delays and bandwidth bottlenecks, and thus restricts the application of *fl* in *iov*. To alleviate channel congestion, existing research has proposed gradient quantization algorithms to reduce communication bandwidth overhead by reducing the transmission of redundant gradient data. However, in the *fl* for *iov*, traditional gradient quantization algorithms face the challenge of trade-off between compression rate and global model performance. Specifically, the amount of data and the critical information it contains are not balanced across different vehicle nodes, and the use of compression with undifferentiated gradient discard may lead to the loss of important information, which in turn degrades the performance of the global model.

Therefore, in order to solve the channel overload problem due to the large number of gradient data interactions in the *fl* for *iov*, while ensuring the stability of the compressed gradient aggregated model in terms of accuracy and convergence, this paper proposes an entropy-based adaptive gradient quantization (*eaqfed*) algorithm. The *eaqfed* addresses the gradient update process of the local model of the vehicle node, and first calculates the entropy value of the gradient update to measure the importance of the gradient. Second, an objective function is constructed based on the entropy value, and dynamic adjustment of the gradient update is achieved by optimizing this objective function to solve for the adaptive quantization level. The method can assign finer quantization levels to gradients with more critical information, while adopting a higher compression rate for gradients with more redundant information, thus effectively reducing the communication overhead while improving the accuracy and stability of the aggregation model. The main contributions of this paper are as follows:

(1) We propose the entropy-based adaptive gradient quantization algorithm, which adaptively adjusts the gradient quantization level based on the entropy values of different vehicle node model updates, to refine the gradient information for different vehicle nodes.

(2) We verify the performance of the *eaqfed* algorithm on two real datasets with multiple baseline algorithms as a control.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 details the process of entropy-based gradient quantization algorithm. Section 4 validates the performance of the entropy-based gradient quantization algorithm on two real datasets, mnist and fashion-mnist. Section 5 concludes the paper.

2. Relative Work

2.1. Gradient Quantisation Algorithm

In the *fl* for *iov*, vehicle nodes represent the original gradient vectors in a more compact form through gradient quantization before uploading the local model update gradients to the roadside unit for local aggregation, thus

significantly reducing the amount of data during the transmission of gradient vectors. In gradient quantization algorithms, each element in the gradient vector is approximated as a finite set of discrete values, thus reducing the bit-width of the element. The *qsgd* [3] represents the gradient by using a lower number of bits, instead of transmitting a full-precision floating-point number (e.g., a 32-bit floating-point) directly. Specifically, it keeps the gradient unbiased by stochastic rounding (stochastic rounding) while reducing the amount of communication. The *signsgd* [4] is an optimization algorithm that updates the model parameters based on the gradient direction (symbols), and the core idea is to transmit only the symbolic information of the gradient instead of the complete floating-point value, thus greatly reducing the communication cost, but the updating direction may be unstable due to the influence of noise in a distributed environment. However, in a distributed environment, the update direction may be unstable due to the influence of noise. Applying the concept of unbiased estimation, *terngrad* [5] converts the gradient to $\{-1,0,1\}$ via triple-valued quantization, thus greatly reducing communication overhead while maintaining training effectiveness as much as possible. It is suitable for ultra-low communication bandwidth scenarios, but the convergence speed may be affected due to the large information loss.

3. Entropy-Based Adaptive Gradient Quantization Algorithm

3.1. Design Quantization Level

In the *fl* for *iov*, on the t-th round of communication between vehicle nodes and roadside units, vehicle node k receives the global model parameters w_c^{t-1} from the roadside unit u and initialises the local model with local model parameters w_k^t. Subsequently the local model is trained to utilise the traffic data $data_k$ collected by the vehicle node, which results in an updated gradient vector v_k^t:

$$v_k^t = [v_{k,1}^t, v_{k,2}^t, \dots, v_{k,d}^t] \in \mathbb{R}^d \tag{1}$$

Where d denotes the gradient vector dimension and $v_{k,d}$ denotes the d-th gradient element in the gradient vector. Let the largest element in the gradient vector v_k^t be $v_{\max}$,and the smallest element be $v_{\min}$,. Thus, the range of gradient values represented by range=$v_{\max} - v_{\min}$.

First, the range is divided into n equal spaced intervals (bins) and the size of each bin is calculated as follows:

$$bin = \frac{|v_{\max} - v_{\min}|}{n} \tag{2}$$

Where N is a hyperparameter that determines the fineness of the division.

Second, the number of gradients in each bin_i, denoted as h_i, and the probability distribution of h_i in each bin is determined as follows:

$$p_i = \frac{h_i}{\sum_j^n h_j} \tag{3}$$

Where, p_i represents the probability that a gradient value falls into bin_i.

Then, the entropy of each bin_i gradient is computed using the standard information entropy formula:

$$e_j = -\sum_{i=1}^{N} p_i \log p_i \tag{4}$$

Where, e_j represents the entropy of the current layer's gradient. A higher entropy value indicates a greater amount of gradient information.

Finally, the hyperparameter 2^r is set, and the constraint range l is defined within $[0,6]$, resulting in:

$$l = \min(6, \max(0, \lceil e/2^r \rceil)). \tag{5}$$

This ensures that l varies within the range $[0,6]$. When the entropy value is high, the gradient contains more information, so the quantization levels increase to retain more information. When the entropy value is low, the gradient contains less information, so the quantization levels decrease to improve the compression ratio.

3.2. Adaptive Gradient Quantization

The quantization level l obtained in the previous section is substituted into the quantization function expression, resulting in:

$$q_l(v_k^t) \triangleq [q_l(v_{k,1}^t), q_l(v_{k,2}^t), \dots, q_l(v_{k,d}^t)] \tag{6}$$

Where, L_i^k represents the quantization level, and $q_l(v_{k,d}^t)$ represents the quantized value of the gradient $v_{k,d}^t$ based on the quantization level l. The hierarchical quantization compression of vehicle nodes k is expressed using the following formula:

$$q_l(v_{k,d}^t) = \|v_i^t\| \cdot sign(v_{k,d}^t) \cdot \xi_k(s) \tag{7}$$

Where, $\|v_i^k\|$ is the norm of the gradient vector v_i^t, which serves as the scaling factor. The function $sign(\cdot)$ represents the sign operation. $s \in \{1,2,\cdots,l\}$ denotes the number of quantization levels, and $\xi_i(s)$ is a random variable following a Bernoulli distribution. After quantization and compression, the probability that the gradient vector of vehicle node k takes the value $l_{s(r)}$ is $1 - \xi(r_d)$, while the probability of taking the value $l_{s(r)+1}$ is $\xi(r_d)$. r represents the normalized coordinates of the gradient vector.

4. Experiments

This section verifies the quantization performance of the *eaqfed* algorithm. Specifically, we compare the *eaqfed* algorithm in *fl* for *iov* with baseline algorithms such as *qsgd*, *signsgd*, and *terngrad* on the mnist and fmnist datasets. These algorithms compress the gradient updates for different rounds to get the accuracy of the compressed aggregated model as well as the loss during training, and model accuracy curves are used to compare the convergence speed of the different algorithms.

4.1. Experiment Setup

This experiment builds a three-layer federated learning simulation model for *iov*, consisting of one cloud server, two roadside units, and six vehicle nodes. Each vehicle node is equipped with a resnet-18 model, and a total of 50 communication rounds are conducted to verify the effectiveness of the *eaqfed* algorithm on real datasets.Datasets: the mnist dataset contains 10 classes of handwritten digits ranging from 0 to 9. The fashion-mnist dataset includes 10 categories of clothing items. These two datasets are mainly used to evaluate the performance of deep learning models in image classification tasks.

4.2. Experimental Results

During the first 50 communication rounds, the accuracy of the aggregated model on the mnist and fashin-mnist datasets is shown in Fig. 1 and Fig. 2, respectively.

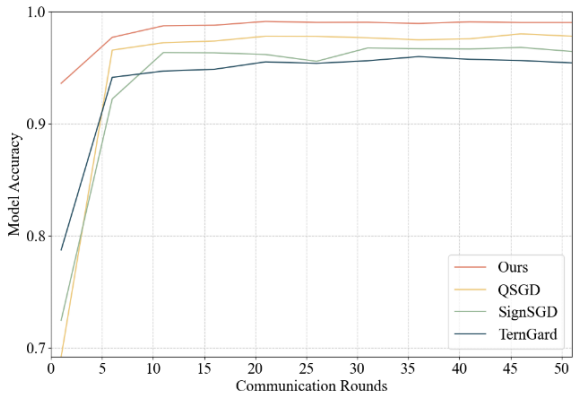

Figure 1. Communication and model accuray in mnist

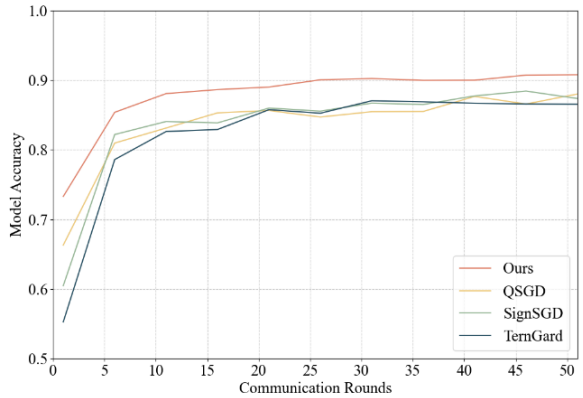

Figure 2. Communication and Model accuracy in fashion-mnist

In **Figure** , it can be observed that the *eaqfed* algorithm maintains the highest model accuracy throughout the training process. It reaches over 95% accuracy within the first 10 rounds and achieves a final accuracy close to 99%. In contrast, *qsgd* shows a better convergence trend, but its final model accuracy is slightly lower than that of *eaqfed*. The *signsgd* and *terngrad* suffer from a greater loss of gradient information, resulting in a lower accuracy of their final models. In **Figure** , the *eaqfed* algorithm consistently maintains a high model accuracy throughout the training process. Compared to other methods, it converges faster and achieves a final accuracy superior to other gradient compression methods. Within 10 communication rounds, *eaqfed* reaches approximately 85% accuracy, and in the subsequent training, it steadily improves, ultimately approaching 90%.

5. Conclusion

To reduce the communication overhead in the *fl* for *iov* while maintaining high accuracy of the aggregation model, this paper proposes an entropy-based adaptive gradient quantization (*eaqfed*) algorithm. The *eaqfed* computes the entropy of gradient updates from different vehicle nodes and assigns them different quantization levels, adaptively adjusting the quantization level based on the redundancy of the data and improving the stability of model convergence.

References

[1] Duan W, Gu J, Wen M, et al. Emerging technologies for 5G-IoV networks: applications, trends and opportunities. IEEE Network. 2020, Vol. 34 (No. 5), p. 283-289.

[2] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics. 2017, Vol. 54 (No. 20), p. 1273-1282.

[3] Alistarh D, Grubic D, Li J, et al. QSGD: Communication-efficient SGD via gradient quantization and encoding. Advances in neural information processing systems. Long Beach, 2017.

[4] Bernstein J, Wang Y X, Azizzadenesheli K, et al. signSGD: Compressed optimisation for non-convex problems. International Conference on Machine Learning. Stockholm, 2018.

[5] Wen W, Xu C, Yan F, et al. Terngrad: Ternary gradients to reduce communication in distributed deep learning. Advances in neural information processing systems. Long Beach, 2017.