Poster: Clipped Quantization and Huffman Coding for Efficient Secure Transfer in Federated Learning

Seung-Ho Lim
Division of Computer Engineering
Hankuk University of Foreign Studies
Yongin, Korea
slim@hufs.ac.kr

Min Choi
Sch. of Information and Communication Engineering, Chungbuk National University
Cheongju, Korea
mchoi@chungbuk.ac.kr

Ki-Woong Park
Dept. of Computer and Information Security, Sejong University
Seoul, Korea
woongbak@sejong.ac.kr

***Abstract*—Federated Learning(FL) has become an emerging method that trains private data by distributed learning of shared parameter, however, it has high communication overhead and is still exposed to attack on the model parameters. To minimize the communication overhead of federated learning while preserving its accuracy and security, we considered a combination of techniques with gradient quantization, clipping, and Huffman coding. Our system produces reduced parameters through quantization and clipping, then encodes the parameters through Huffman coding, which further increases the compression ratio as well as security of the parameters to be transmitted. Preliminary results identify that the scheme can significantly reduce the amount of transferring while preserving accuracy and security.**

***Index Terms*—Federated Learning, Quantization, Clipping, Huffman Coding**

I. Introduction

A federated learning(FL) has been proposed to solve the problems of personal data use causing privacy issues, in which it is a learning model that cooperates by sharing only learned parameters rather than directly sharing data stored in local clients [1], [2]. FL increases the accuracy by performing learning through repeated up-down of the model parameters between the clients and server, so there cane be overhead in data communication in spite of only parameters are shared. Moreover, exposure of parameters still presents security vulnerabilities and makes it a target for attacks such as gradient inversion attacks [3], [4].

Various studies have been conducted on federated learning system such as model compression, pruning, quantization, and optimization of transmissions [5]–[8]. However, there is still a need for advanced methods to enhance security while reducing of communication. Quantization is a representative method of compressing a model for reducing parameter size. Gradient quantization in FL minimizes loss of accuracy even with a small number of bits. Many existing researches have focused to reduce quantization or compression with coding bits while maintaining accuracy as much as possible [6]–[9], or have studied quantization methods considering various heterogeneous devices [10]. Some research to increase the security of transmitted parameters also increased the amount of transmitted data according to encryption techniques [5].

Focusing on the distribution of quantized gradients, we consider ways to enable secure communication while reducing the amount of communication. To reduce the communication overhead of federated learning while preserving its accuracy and security, we developed a combination of several techniques with gradient quantization, clipping, and Huffman coding in federated learning system. We produce reduced parameters through gradient quantization and clipping [5], then, encode the parameters through Huffman coding [9], which further increases the compression ratio as well as security of the parameters to be transmitted. Preliminary results identify that those scheme shows the desired effect.

II. Clipped Quantization and Huffman Coding

Since the amount of change in parameters is not large each time they are learned in an iterative process, gradient quantization is performed when quantizing parameters in federated learning to reduce the amount of transmission. Although transmission volume is reduced through gradient quantization, it may be vulnerable to security threats such as inversion attacks due to the consistent size and exposure of each parameter size. Moreover, the fixed size of the parameters is inefficient when looking at the distribution of the parameters. In order to further reduce the size of gradient quantized parameters as well as increase security at the same time, it is necessary to apply a data encoding technique based on the distribution of parameters.

In general, the distribution of quantized gradients has near Gaussian distribution with a high ratio in the middle, while the ratio at both ends is extremely low. Based on this distribution, we apply clipping for quantization to create the minimum valid bits representation required for transmission [7]. For example, when performing 4-bit quantization, clipping 2 bits at both ends can reduce the total data by 75%. In order to reduce the actual amount of transmission for clipped data, we apply Huffman coding to the clipped data. It not only reduces the bitstream of the transmitted code, but also increases security against parameter attacks because it also causes data distortion through encoding.

Figure 1 shows the developed federated learning system with clipped quantization and Huffman coding applied. As shown in the figure, the quantized gradients go through a clipping process and it removes several bits at both ends with the pre-defined clipping rate, which results in the number

2837-4827/24/$31.00 ©2024 IEEE
DOI 10.1109/SEC62691.2024.00065

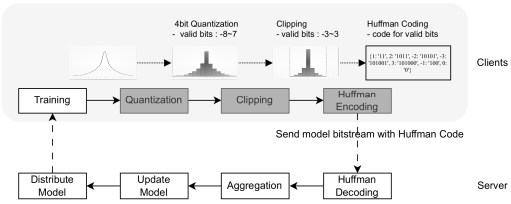

Fig. 1. Overview of FL with clipped quantization and Huffman coding

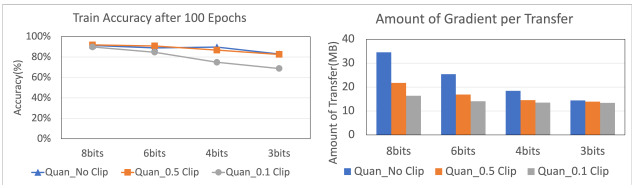

Fig. 2. Experimental Results for Clipped Quantization and Huffman Coding

of valid bits decreases significantly compared to the total quantization bits. In addition, the data has different ratios for amount between valid bits, which means high rates in the middle and low rates at both sides. When Huffman coding is applied to these, the data representation is distorted according to the distribution of valid bits. That is, the expression for bits with a high ratio is simplified, and the bit representation for bits with a low ratio is increased. In addition, the encoded data stream size is much reduced. The client transmits the generated bitstream to the server along with the Huffman code tree. The server performs decoding jobs for the encoded bitstream from each client and the Huffman code tree. Then, it performs aggregation according to federated learning algorithm, and the updated model is distributed to clients.

III. Experimental Results

We have implemented the clipped quantization and Huffman coding scheme on the existing open federated learning platform [5], in which it supports interfaces for gradient quantization with clipping and learns parameters with pre-set quantization level and clipping ratio. With the implemented system, we did preliminary experiments on FL with Resnet-18 deep learning model using CIFAR-10 dataset, which used RMSprop optimizer having 10^{-6} decay for optimizer. We performed training for 100 epochs to the Resnet-18 model on 10 clients by changing the quantization level with bits with 8, 6, 4, and 3 bits, and with the clipping level of 0, 0.5, and 0.1 for each quantization level. For clipping level, 0-level means no clipping, 0.5-level means 50% of parameters from center is remained, while 0.1 means only 10% from center is remained.

The experimental results are shown in Figure 2, where the left side represents the accuracy and the right side represents the amount of transmitted data. As shown in Figure 2, if we increases clipping threshold level, the amount of data transmitted through Huffman coding greatly decreases. While the accuracy is not much degraded for 0.5 clipping level, too much clipping significantly reduces accuracy such as 0.1 level. When comparing 0.5 clipping with no clipping, we identify that there is little difference in accuracy while the amount of data transmission is greatly reduced. It means that the amount of transmission can be greatly reduced while preserving accuracy through proper clipping and Huffman coding at the same time. The Huffman coding combined with appropriate bit level quantization and clipping can maintain optimal accuracy with minimal communication overhead as well as security with distorted data by encoding.

IV. Conclusion and Further Work

In federated learning, the vulnerabilities due to transmission volume and parameter attacks should be improved. We developed the clipped quantization and Huffman coding-based federated learning system. The experimental results shows that the transmission amount of gradient parameters could be reduced with little decrease in accuracy. For the further work, we will perform an evaluation for the gradient inversion attack, and also analyze the execution time overhead of Huffman coding. Based on the the experimental results, we will conduct further research on how to find the optimal clipping levels for each client in federated learning, compression methods in comparison with previous works as well as seek the feasibility of Huffman coding against attacks.

References

[1] akubKonečný, H.BrendanMcMahan, DanielRamage, and PeterRichtárik, "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," arXiv:1610.02527, 2016.

[2] Andrew Hard and et al. "Federated Learning for Mobile Keyboard Prediction,", in arXiv:1811.03604, 2019.

[3] PretomRoyOvi, EmonDey, NirmalyaRoy, and AryyaGangopadhyay, "Mixed Quantization Enabled Federated Learning to Tackle Gradient Inversion Attacks," In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 5046–505., 2023.

[4] Ligeng Zhu, Zhijian Liu, and Song Han, "Deep Leakage from Gradients," In Advances in Neural Information Processing Systems, Vol. 32. Curran Associates, Inc., 2019.

[5] Chengliang Zhang, SuyiLi, JunzheXia, WeiWang, FengYan, and Yang Liu., "BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning," In 2020 USENIX Annual Technical Conference (USENIX ATC 20). pp. 493–506, 2020.

[6] DÉFOSSEZ, Alexandre; ADI, Yossi; SYNNAEVE, Gabriel, "Differentiable model compression via pseudo quantization noise," arXiv preprint arXiv:2104.09987, 2021.

[7] WeiWen, CongXu, FengYan, ChunpengWu, YandanWang, YiranChen, and Hai Li., "Terngrad: Ternary gradients to reduce communication in distributed deep learning," Advances in neural information processing systems 30, 2017.

[8] Reisizadeh, Amirhossein, et al. "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," International conference on artificial intelligence and statistics. PMLR, 2020.

[9] Song Han, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 4th International Conference on Learning Representations(ICLR), 2016.

[10] Gupta, Kartik, et al. "Quantization robust federated learning for efficient inference on heterogeneous devices," in Transactions on Machine Learning Research, 2023.