2ND WORKSHOP
"MACHINE LEARNING & NETWORKING" (MaLeNe)
PROCEEDINGS

SEPTEMBER 4, 2023

CO-LOCATED WITH
THE 5TH INTERNATIONAL CONFERENCE ON NETWORKED SYSTEMS (NETSYS 2023)
POTSDAM, GERMANY

Parameter Prioritization for Efficient Transmission of Neural Networks in Small Satellite Applications

Olga Kondrateva
Technical University of Darmstadt
olga.kondrateva@kom.tu-darmstadt.de

Stefan Dietzel
Merantix Momentum GmbH
stefan.dietzel@merantix.com

Ansgar Lößer, Björn Scheuermann
Technical University of Darmstadt
{ansgar.loesser, scheuermann}@kom.tu-darmstadt.de

***Abstract*—Low-earth-orbit (LEO) satellites can be used for cost-effective Earth observation missions. Onboard processing using machine learning (ML) approaches is often proposed to reduce the amount of data transmitted back to Earth. However, the combination of LEO satellites and ML brings unique communication challenges, as requirements – and therefore ML models – often change throughout the lifetime of a satellite mission. In this paper, we propose a novel communication protocol that deals with model updates efficiently by providing incremental updates with low communication overhead.**

I. Introduction

Small, low-Earth-orbit (LEO) satellites allow us to deploy satellite missions more quickly and cost effectively. In particular, the CubeSat standard became popular due to the availability of off-the-shelf components [1]. Often, the amount of data acquired by the LEO satellites is too large to transmit everything to Earth [2] during their short, unreliable communication windows. To intelligently filter the most relevant information, machine learning have gained rising attention in the satellite community. But their deployment raises new communication issues in the upstream direction: how can machine learning models be updated efficiently?

We propose a novel communication protocol, which allows for efficient incremental model updates. Due to its large number of parameters, the updated model's transmission likely requires use of multiple communication windows. Yet, the new model must be used as soon as possible in order to benefit from the updated accuracy or adapt to new classification tasks quickly. Therefore, our communication protocol prioritizes the most important model weights in transmission. We use a space-efficient data structure to convey priority classes to the satellite with low communication overhead. Once its most important weights are received, they can be used to construct an approximation of the updated model. The approximation is then used immediately, and it is improved incrementally until all updated model weights are available.

Evaluation results show that our approach considerably outperforms the baseline and performs similar to an ideal update protocol while incurring significantly less overhead.

Next, we introduce our approach in Section II and evaluate it in Section III. Section IV concludes the paper.

The full version of this paper has previously been published at the IEEE MedComNet 2023 conference: https://www.medcomnet.org.
This work has been funded by the Federal Ministry of Education and Research of Germany in the project "Open6GHub" (16KISK014).
This work received funding from the German Research Foundation (DFG), CRC 1404: *FONDA: Foundations of Workflows for Large-Scale Scientific Data Analysis.*
This work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

II. Efficient Incremental Weight Transmission

Recall that a machine learning model mainly consists of a description of its structure and a number of weights that describe how these neurons fire. We consider the model structure to be known in advance, as standard structures are often used for common tasks. Therefore, each model update can be considered as transmitting a list of new model weights.

We design our protocol such that the newly updated model can be approximated quickly using all weights received up to a certain point in time with the remainder of the weights all set to zero. Two questions arise in this context: (1) What is the *right* order of model weights. (2) How can we communicate the prioritized order of model weights efficiently?

A. Weight order representation

Our approach depends on prioritizing key parameters. To determine importance, we draw on ideas from model pruning, where less crucial neural network components are removed. Various criteria, like the magnitude criterion [3], L1 and L2 norms [4], and gradient magnitude [5] have been used to measure importance. In our work, we utilize the absolute magnitude criterion, ranking parameters by their absolute value, and deeming those with lower values less important.

Next, we design a compact representation of the model weights' order, using a lossy permutation compression approach. To this end, we adopt *sorting subsequences,* a straightforward yet shown to be optimal compression scheme [6]. Rather than encoding the exact priority order for *each* model weight, we subdivide the list of weights into k *groups* of decreasing priority. This approximate grouping reduces the size of additional communication overhead to $n \log(k)$ bits.

B. Transmission protocol

Figure 1 shows our approach for a simplified model with six weights. As we assume the model structure to be known, we can represent a model by a flattened array of weights W.

Next, we perform a number of initialization steps on the ground station before transmission starts. We calculate a permutation P that prioritizes each model weight by its absolute value (Step 2). Then, we divide the array P into k groups of length m. First, the vector Q that maps each weight index to its priority group is transmitted to the satellite (Step 3), followed by all weights of the priority group 0. Within the priority group, the weight with the smallest index i is transmitted first, followed by the second-smallest, and so forth (Steps $4, \ldots, n$). Therefore, the order within the group does not need to be communicated to the satellite but can be inferred from the index structure Q. When all weights of priority group 0 have been received, the process continues with priority group 1, and so forth until all weights have been

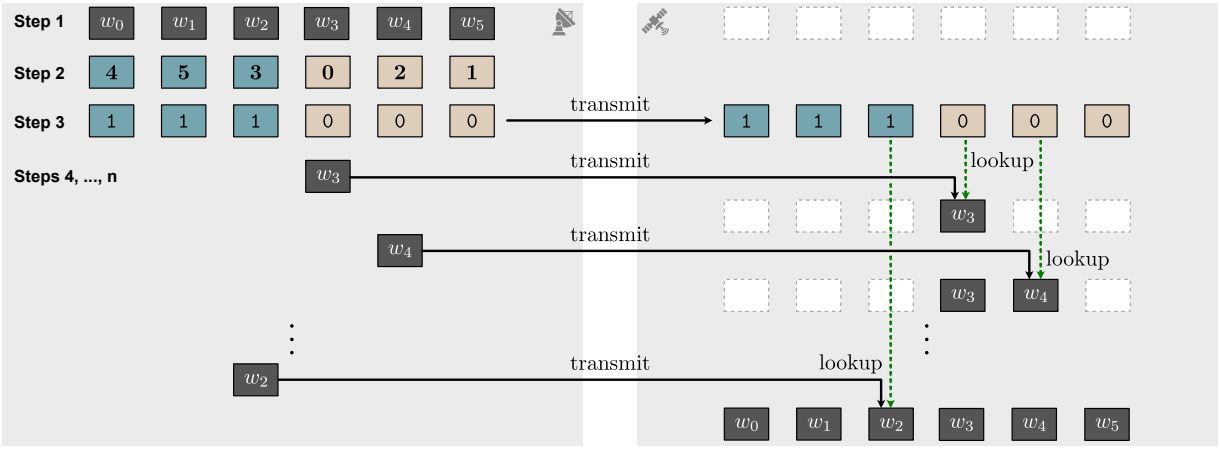

Fig. 1. Overview of the proposed communication scheme for an example model with six weights where $k = 2$ and $m = 3$.

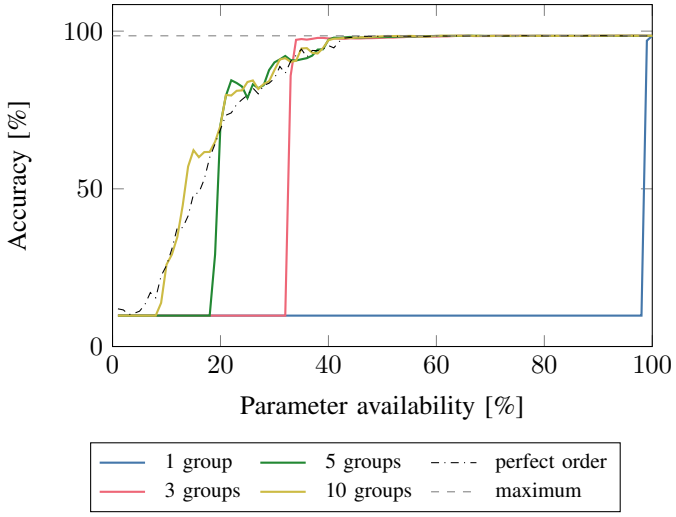

Fig. 2. Accuracy improvements depending on percentage of available model weights for MNIST trained on LeNet5.

received. In the example, the weights are transmitted in the order w_3, w_4, w_5 (group 0) followed by w_0, w_1, w_2 (group 1).

At the satellite, the updated model is used immediately by setting all unknown model weights to 0 as indicated by the dashed white boxes in Figure 1. Whenever more weights have been received, the model is incrementally updated on the satellite until the fully updated model is available.

III. Evaluation

To evaluate our approach, we use the LeNet5 [7] model trained on the MNIST [8] dataset. We compare our approach against a baseline and an optimal approach. The baseline is the naïve approach that puts all parameters into a single priority group, which achieves no prioritization. The optimal approach assumes that the weights are transmitted in the optimal order, i. e., according to their absolute value, highest first. In this case, the position of each weight in the model's structure needs to be communicated to reconstruct the model on the satellite. To isolate our mechanism's influence on accuracy, we evaluate it independent of communication effects, simply assuming that weights are available in a certain order.

Figure 2 shows our evaluation results. The x-axis shows the percentage of weights transmitted so far. All other weights are assumed to be 0. The y-axis shows the corresponding classification accuracy of the incrementally updated model. We compare different numbers of priority groups to assess how quickly they achieve good accuracies. It becomes clear that putting all parameters into one priority group does not allow for partial updates, since all parameters are required to achieve a meaningful accuracy level. It also can be seen that our approach allows to find a good tradeoff between the achieved accuracy and the amount of additional data that has to be transmitted. In addition, the results show that even the optimal parameter order gets outperformed, which indicates that the order of weights chosen (by absolute value) is not the only influence factor for model accuracy. In summary, the proposed approach considerably improves the accuracy when compared to the baseline and even the optimal approach.

IV. Conclusion

Performing efficient, incremental updates of machine learning models on satellites is a problem that has often been neglected. However, it is imperative to enable widespread use of LEO satellites despite changing classification requirements during satellite operation and short contact times with base stations. We have proposed a simple but effective mechanism to perform incremental model updates based on prioritizing weights into groups. Using this group-based ordering, we achieve significantly faster improvements in classification accuracy while keeping communication overhead to convey the prioritization order low.

References

[1] K. Woellert, P. Ehrenfreund, A. J. Ricco, and H. Hertzfeld, "Cubesats: Cost-effective science and technology platforms for emerging and developing nations," *Advances in Space Research*, vol. 47, no. 4, 2011.

[2] G. Furano, A. Tavoularis, and M. Rovatti, "Ai in space: applications examples and challenges," in *2020 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 2020.

[3] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," *arXiv preprint arXiv:1506.02626*, 2015.

[4] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv:1608.08710*, 2017.

[5] N. Lee, T. Ajanthan, and P. H. S. Torr, "Snip: Single-shot network pruning based on connection sensitivity," *arXiv:1810.02340*, 2018.

[6] D. Wang, A. Mazumdar, and G. W. Wornell, "Compression in the space of permutations," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, 2015.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, 1998.

[8] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012.