Communication Optimization in Large Scale Federated Learning using Autoencoder Compressed Weight Updates

Srikanth Chandar, Pravin Chandran, Raghavendra Bhat & Avinash Chakravarthi

Abstract

The communication of model weight updates in Federated Learning (FL) comes with significant network bandwidth costs. We propose a mechanism of compressing the weight updates using Autoencoders (AE), where weights are encoded before transfer at "Collaborators" and decoded back at "Aggregator". The model parameters corresponding to the encoder and decoder are learnt using a pre-pass training inside Collaborators. This model weights driven and orthogonal AE based weight compression technique serves as a complementary approach to traditional compression methods in a large-scale FL. It not only achieves compression ratios ranging from 500x to 1720x and beyond, but also is flexible as it can be customized based on the accuracy requirements, computational capacity, and other requirements of the given FL setup.

Keywords: Federated Learning, Communication Efficient, Autoencoder, Compression.

Introduction

The weight updates during Federated Learning are non-isolated events and there is a relation and interlinking between the parameters. An AE network can find such patterns hidden in the data and reduce the representation to a lower-dimensional feature size. This paper investigates and shows that an AE can exploit this fact of dependence between parameters in a weight update, and thus learn the encoding of the weight update, and also replicate it in a "learnable" manner.

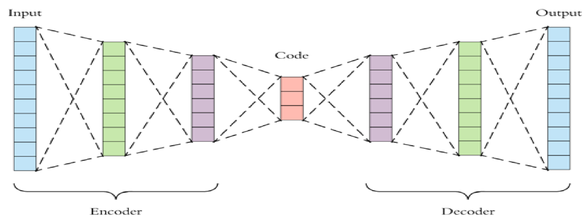


Figure 1: Autoencoder model

Architecture- Prepass

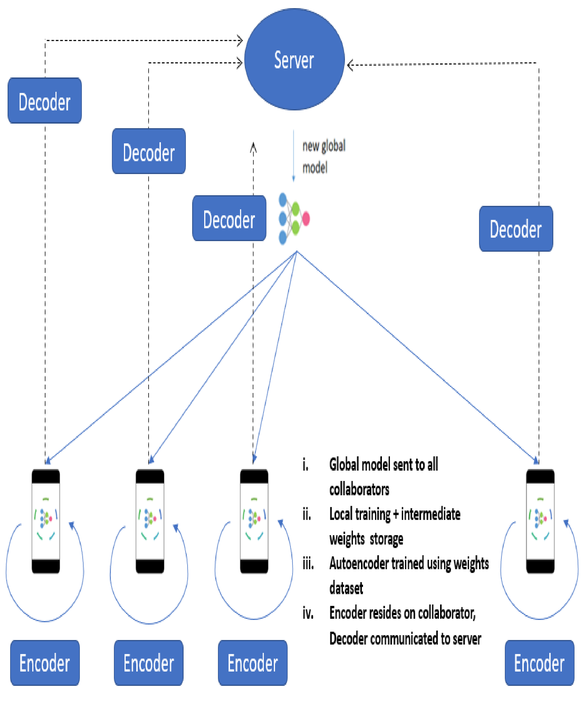


Figure 2: Prepass Round

Architecture- FL

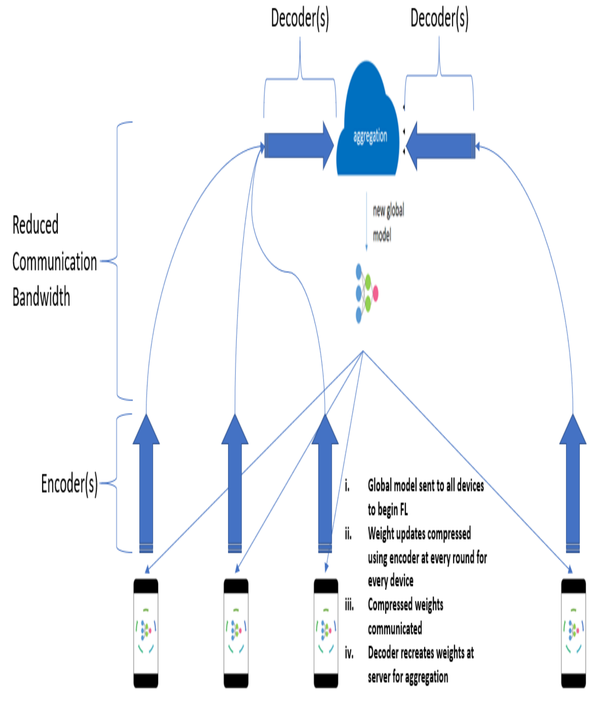
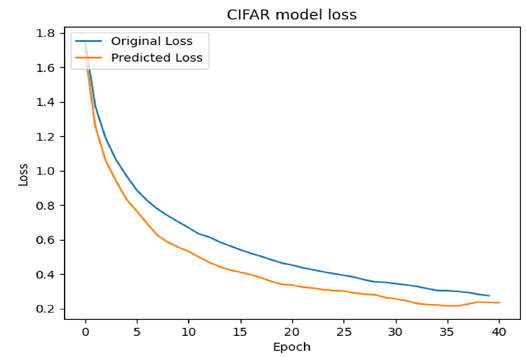
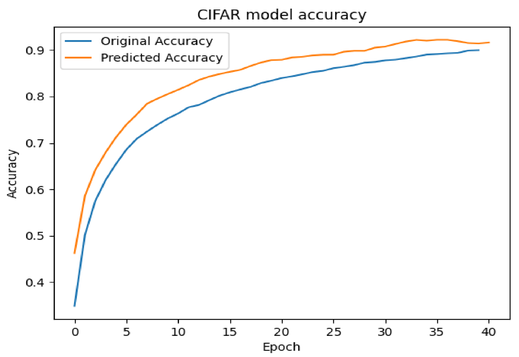
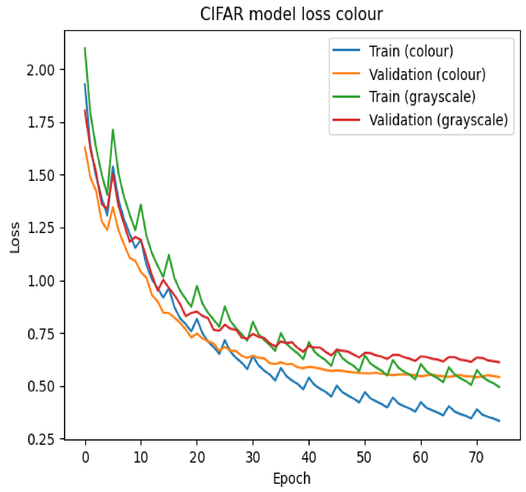


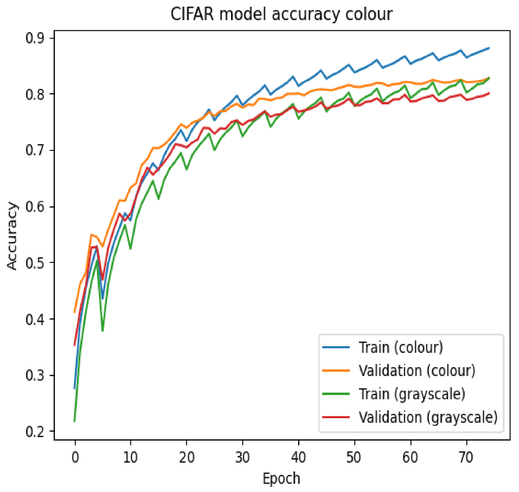
Figure 3: FL with AE compression

Results









Conclusion

- Largest compression (lossy) ratios achieved: 2000x
- A trade-off-based analysis is needed for AE-based compression. Since there is overhead of decoder communication in prepass round, this setup is more useful for large scale FL (large no. of collaborators/communication rounds)
- AE compression for MNIST and CIFAR classifier is tested in the FL setup.
- Compression ratio is dynamically set based on computational capacity, accuracy requirement, etc. This method is orthogonal and can complement certain other compression mechanisms.

Future Direction

Future work can explore the use of convolutional layers in auto-encoder to significantly reduce AE parameters. Also, impact of non-iidness in training data distribution can be studied to further stress-test the generalizability of this method.

Key References

[1] Hongyi Wang et al. Federated learning with matched averaging.

[2] Alexander Frickenstein et al. Alf: autoencoder-based low-rank filter-sharing for efficient convolutional neural networks.