AQUILA: COMMUNICATION EFFICIENT FEDERATED LEARNING WITH ADAPTIVE QUANTIZATION OF LAZILY-AGGREGATED GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The development and deployment of federated learning (FL) have been bottlenecked by the heavy communication overheads of high-dimensional models between the distributed client nodes and the central server. To achieve better error-communication tradeoffs, recent efforts have been conducted to either adaptively reduce the communication frequency by skipping unimportant updates, a.k.a. *lazily-aggregated quantization (LAQ)*, or adapt the quantization bits for each communication. In this paper, we started from the careful investigation on the classical LAQ scheme and theoretically proved that the quantization bit for each communication round in LAQ can be optimized by minimizing the lazy aggregation error. Based on the above theoretical finding, we proposed a unified communication-efficient framework for FL based on **a**daptive **qu**ant**i**zation of **l**azily-**a**ggregated gradients (**AQUILA**), which adaptively adjusts the communication frequency and quantization bit in a synergic way. Specifically, we designed a novel optimal quantization criterion based on gradient updates for lazy aggregation, and adjusted the LAQ strategy to better fit the adaptive quantization. The effectiveness and convergence of the proposed **AQUILA** framework are theoretically verified. The experimental results demonstrate that **AQUILA** can reduce around 50% of communication bits compared to existing methods without accuracy degradation. The proposed **AQUILA** is highly adaptive and compatible to existing FL settings.