Automatic Tuning of Federated Learning Hyper-Parameters from System Perspective

Research Lead: Prasant Mohapatra, Department of Computer Science, UC Davis, pmohapatra@ucdavis.edu

Research Goals

Background

- Federated Learning (FL) as a powerful paradigm to preserve data privacy while enabling Machine Learning (ML) training for various applications.
- FL training results in large system overheads with regards to Computation Time (CompT), Transmission Time (TransT), Computation Load (CompL), and Transmission Load (TransL).
- Applications have different preferences on CompT, TransT, CompL, and TransL. E.g., Anomaly detection is time-sensitive (CompT + TransT); Precision Agriculture is sensitive to energy consumption (ComL + TransL).

Goals

- Automatic FL hyper-parameter tuning algorithm to reduce system overheads of FL training.
- Algorithm respects application's preference on CompT, TransT, CompL, and TransL.
- Algorithm needs to tune hyper-parameters during FL training. No "comeback" is allowed as the FL model keeps training until its final model accuracy. Otherwise, more system overheads are introduced.
- Algorithm is easily integrated to standard FL training settings.

Technical Approach

Problem Formulation

- For two FL hyper-parameter sets S_1 and S_2, we can decide which one is better by the following comparison function $I(S_1, S_2)$

$$I(S_1,S_2) = \alpha \times \frac{t_2 - t_1}{t_1} + \beta \times \frac{q_2 - q_1}{q_1} + \gamma \times \frac{z_2 - z_1}{z_1} + \delta \times \frac{v_2 - v_1}{v_1}$$

where α, β, γ, and δ are application's preference on CompT, CompL, TransT, and TransL. t, q, z, v are the measurements of CompT, CompL, TransT, and TransL.

- If $I(S_1, S_2) < 0$, then S_2 is better than . A set is better if the weighted improvement of some training aspects (e.g., CompT and CompL) is higher than the weighted degradation of the remaining training aspects (e.g., TransT and TransL).
- However, the training overheads for different sets are unknown *a priori*. As a result, directly identifying the optimal hyper-parameters before FL training is impossible.

- <u>We propose an iterative method to optimize the next set of hyper-parameters</u>. Given the current set S_{cur}, the goal is to find a set S_{nxt} that minimizes the following objective function

$$G(S_{nxt}) = \alpha \times \frac{t_{nxt} - t_{cur}}{t_{cur}} + \beta \times \frac{q_{nxt} - q_{cur}}{q_{cur}} + \gamma \times \frac{z_{nxt} - z_{cur}}{z_{cur}} + \delta \times \frac{v_{nxt} - v_{cur}}{v_{cur}}$$

Optimization

- We take the number of selected FL clients M as an example to illustrate the optimization process. That is, we optimize set $S_{nxt} = \{M_{nxt}\}$.
- To find the optimal M_{nxt}, we can take the derivative of $G(S_{nxt})$ over M, obtaining

$$\Delta M = \frac{\partial G(S_{nxt})}{\partial M} = \frac{\alpha}{t_{cur}} \times \frac{\partial t_{nxt}}{\partial M} + \frac{\beta}{q_{cur}} \times \frac{\partial q_{nxt}}{\partial M} + \frac{\gamma}{z_{cur}} \times \frac{\partial z_{nxt}}{\partial M} + \frac{\delta}{v_{cur}} \times \frac{\partial v_{nxt}}{\partial M}$$

- We know that CompT and TransT prefer larger M, and CompL and TransL prefer smaller M. Further, we apply a linear function for approximation. Finally, we can calculate ΔM as

$$\Delta M = \frac{(+1) \times \alpha \times \eta_{t-1} \times |t_{cur} - t_{prv}|}{t_{cur}} + \frac{(+1) \times \beta \times \eta_{q-1} \times |q_{cur} - q_{prv}|}{q_{cur}} + \frac{(-1) \times \gamma \times \eta_{z-1} \times |z_{cur} - z_{prv}|}{z_{cur}} + \frac{(-1) \times \delta \times \eta_{v-1} \times |v_{cur} - v_{prv}|}{v_{cur}}$$

- where η are self-calculated parameters.

- If ΔM > 0, then $M_{nxt} = M_{cur} + 1$; otherwise, $M_{nxt} = M_{cur} - 1$

Evaluation

Setup

- Optimize the number of selected clients M and the number of training passes E, i.e., S = {M, E}
- Dataset: Speech-to-command voice, EMNIST handwriting, and Cifar-100 image
- ML model: ResNet, MLP
- Aggregation methods: FedAvg, FedNova, FedAdagrad
- Compared to fixed M and E of 20

Result

Dataset	Speech-command	EMNIST	Cifar-100
Data Feature	Voice	Handwriting	Image
ML Model	ResNet-10	2-layer MLP	ResNet-10
Performance	-22.48%	-8.48%	-9.33%

Table 1. Our performance for diverse datasets when FedAvg aggregation method is applied

Aggregator	FedAvg	FedNova	FedAdagrad
Performance	-22.48%	-23.53%	-26.75%

Table 2. Our performance for diverse aggregation algorithms.

Research Plan / Next Steps

We plan to explore the following research directions.

- Heterogeneous devices and networking conditions
- Participant selection algorithms
- Adaptive training passes across participants
- More aggregation methods

Publication:

Huanle Zhang, Mi Zhang, Xin Liu, Prasant Mohapatra, and Michael DeLucia. FedTune: Automatic Tuning of Federated Learning Hyper-Parameters from System Perspective, arXiv: 2110.03061

Primary Researchers

Prof. Prasant Mohapatra, UC Davis
Prof. Xin Liu, UC Davis
Dr. Huanle Zhang, UC Davis
Dr. Michael DeLucia, ARL

Collaborations

Prof. Cleotilde Gonzalez, CMU