Communication and Energy Efficient Decentralized Learning Over D2D Networks

Dingzhu Wen

School of Information Science and Technology (SIST)
ShanghaiTech University

wendzh@shanghaitech.edu.cn

May 15, 2024

Overview

Outline

Wireless distributed learning

Distributed learning:

- The current data has distributed characteristics, privacy and security requirements are increasing, especially in special scenarios, such as banks and hospitals, which leads to the problem of information islands.
- Increased performance requirements for intelligent applications in distributed scenarios

Wireless communication:

- Mobile edge computing, D2D communication and other wireless communication technology support
- With the increase in mobile devices, the demand for emerging applications, such as autonomous driving and smart recommend systems, is also increasing.

Background: Challenges

Wireless Distributed Learning Performance Evaluation Challenges:

- Distributed training framework: Data distribution of different terminals, model aggregation frequency, model aggregation accuracy, etc.
- Wireless network: Wireless channel noise, communication bandwidth allocation, etc.

Model parameter optimization challenges:

- The spectrum resources of wireless communication network are limited, and the wireless channel is complex and changeable.
- The computing power of the mobile terminal is limited, and the local model calculation will also reduce the efficiency of model training.

Wireless Algorithm Design Problems:

- E.g., user selection algorithm, wireless resource allocation algorithm, link selection algorithm.

Motivation

Motivation:

- In order to realize wireless distributed learning more conveniently, the decentralized learning system is one of the options.
- Due to the limited wireless resources, D2D link selection and wireless resource allocation have a great impact on the model training delay, and also determine the model convergence rate.

Contribution:

- Investigate the framework of joint computing power adjustment, wireless resource allocation, aggregation weight adaptation, and link selection for decentralized learning over D2D networks
- An optimization problem is proposed to minimize the total learning cost by optimizing the computing power, wireless resource allocation, aggregation weight, and link selection.
- Demonstrate the effectiveness of the proposed algorithm by extensive experiments and show that link selection can reduce the learning cost

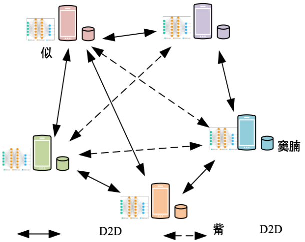
Outline

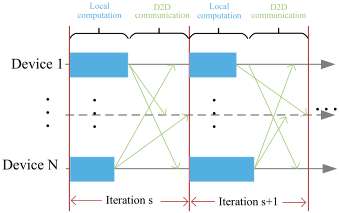
System model

N: the number of local devices, a_{ij}: the connection between device i and device j.

Wireless communication model: The achievable data rate of device i from the link between devices i and j is

$$r_{i,j} = \ell_{i,j} \log_2 \left(1 + \frac{p_{i,j} h_{i,j}}{N_0} \right)$$





System model: Decentralized learning model

Local model calculation:

$$\omega'_{(s+1,i)} = \omega_{(s,i)} - \alpha \mathbf{g}_{(s,i)}$$

Local model aggregation:

$$\omega_{(s+1,i)} = \sum_{i \in \mathcal{N}} x_{i,j} \omega'_{(s+1,i)}$$

Relation between $x_{i,j}$ and $a_{i,j}$:

$$\begin{cases} x_{i,j} = 0; & a_{i,j} = 0 \\ 0 < x_{i,j} \leq 1; & a_{i,j} = 1. \end{cases}$$

Outline

Performance Analysis

If the D2D network requires a maximum number of times of training to converge to the correct rate requirement, the upper boundary of the convergence step can be expressed:

$$S = \frac{\varphi^{\frac{1}{\epsilon}}}{1 - \rho(X)^2}.$$

$\rho(X)$ is the second largest eigenvalue of the network and can indicate the degree of connectivity of the network

$$\rho(X) = \max\{|\lambda_2(X)|, |\lambda_N(X)|\}.$$

In order to ensure that the model training can converge, the D2D network should remain connected, i.e., $\rho(X) < 1$.

Performance Analysis

Per-iteration Learning Latency:

- Local model computation: Denote b as the batchsize, d^C as the CPU cycle required for one training data calculation including both forward and back propagation, and f_i as the computing power of device i, latency can be expressed as

$$t_i^{comp} = \frac{bd^C}{f_i}$$

- Model transmission: Denote M as the volume of model parameters. Then, the latency for model transmission from device i to j can be expressed as

$$t_{i,j}^{comm} = \frac{M}{r_{i,j}}$$

Thus, the per-iteration learning latency can be written as

$$t = \max_{i,j \in \mathcal{N}, i \neq j} a_{i,j} \left(t_i^{comp} + t_{i,j}^{comm} \right)$$

Performance Analysis

Per-iteration Energy Consumption:

- Local model computation: With the latency analysis on the local model computation, the energy consumption of device i on this part can be expressed as
$$e^{comp} = \kappa f_i^2 b d^C,$$
where κ is a coefficient determined by the chip architecture.
- Model transmission: the energy consumption of device i is
$$e_i^{comm} = \sum_{j,j\neq i} a_{i,j}\ell_{i,j}p_{i,j}\frac{M}{r_{i,j}}$$

Thus, the per-iteration energy consumption can be written as
$$e = \sum_i e_i^{comm} + e_i^{comp}.$$

The total learning cost can be expressed as
$$C = (\beta e + (1-\beta)t)\frac{\varphi \cdot \frac{1}{\epsilon}}{1-\rho(X)^2}.$$

Outline

Problem formulation

Analysis:

- D2D link selection plays an important role in learning performance.
  - The number of selected D2D links affects both the learning cost and convergence rate.
  - Due to the dynamic wireless channels, the appropriate link selection strategy will reduce the learning latency and energy consumption.
- The aggregation weight affects the convergence rate by the connectivity of the D2D network.
- Computing power and wireless resource allocation have an impact on both the per-iteration learning latency and per-iteration energy consumption

Therefore, to improve both the communication and energy consumption efficiencies of the D2D-assisted decentralized learning, the link selection, aggregation weight, computing power, and wireless resource allocation should be jointly optimized to minimize the total learning cost.

Problem formulation

The optimization problem is formulated as

$$
\begin{aligned}
\min_{\{a_{i,j},\ell_{i,j},f_i,x_{i,j},p_{i,j}\}} \quad & (\beta e + (1-\beta)t)\frac{\varphi \cdot \frac{1}{\epsilon}}{1-\rho(X)^2} \\
\text{subject to} \quad & \sum_{i,j\in\mathcal{N}} \ell_{i,j} \leq B, \\
& f_i^{\min} \leq f_i \leq f_i^{\max}, i \in \mathcal{N}, \\
& \sum_{i\in\mathcal{N}, i\neq j} \ell_{i,j} p_{i,j} \leq P_i^{\max}, \\
& a_{i,j} = a_{j,i} \in \{0,1\}, \\
& \mathbf{X} = \mathbf{X}^\top, \\
& \mathbf{X}\mathbf{1} = \mathbf{1} \\
& 0 \leq x_{i,j} \leq a_{i,j}, i,j \in \mathcal{N}, i \neq j, \\
& \rho(\mathbf{X}) < 1,
\end{aligned}
$$

Outline

Resource allocation and aggregation weight optimization

Optimal Computing Power, Bandwidth Allocation:

$$\min_{\{\ell_{i,j}, f_i, p_{i,j}, t\}} \beta e + (1-\beta)t, \tag{1}$$

Power Allocation:

$$\begin{aligned}\min_{\{p'_{i,j}, t\}} \beta(&\sum_i \kappa f_i^2 bd^C \\ &+ \sum_i \sum_j a_{i,j} \frac{MN_0}{h_{i,j}} p'_{i,j} (2^{\frac{1}{p'_{i,j}}} - 1)) \\ &+ (1-\beta)t\end{aligned} \tag{2}$$

$$\text{subject to} a_{i,j}\left(\frac{bd^C}{f_i} + \frac{M}{\ell_{i,j}} p'_{i,j}\right) \le t,$$

$$\sum_j \ell_{i,j} \frac{N_0}{h_{i,j}} (2^{\frac{1}{p'_{i,j}}} - 1) \le P_i^{\max}$$

Optimal Aggregation Weight:

$$\min_{i,j} \frac{\varphi \cdot \frac{1}{\epsilon}}{1 - \rho(\mathbf{X})^2},$$

which is equivalent to

$$\min_{x_{i,j}} \|\mathbf{X} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\|_2$$

Algorithm

The Alternating Optimization Algorithm for Computing Power and Wireless Resource Allocation Optimization:

Input: Link selection $a_{i,j}$, channel state information $h_{i,j}$, transmission power constraints P_i^{max}, local computing power constraints $\{f_i^{min}, f_i^{max}\}$, and other system parameters.
Output: Computing power f_i^*, bandwidth allocation $\ell_{i,j}^*$, and transmission power allocation $p_{i,j}^*$.
Initialize power allocation $p_{i,j}$.
repeat
  Under $p_{i,j}$, obtain f_i and $\ell_{i,j}$ by solving the problem in (1)
  Under f_i and $\ell_{i,j}$, obtain $p_{i,j}$ by solving the problem in (2)
until Convergence
Obtain the optimal f_i^*, $\ell_{i,j}^*$, and $p_{i,j}^*$

Key Conclusions

The optimal f_i and $\ell_{i,j}$:

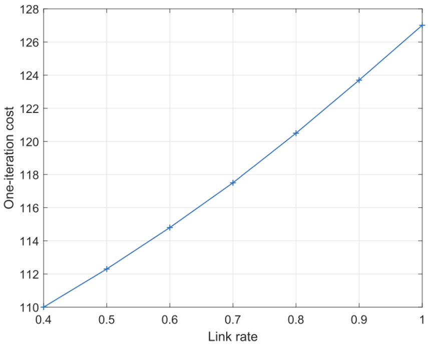
$$f_i^* = \begin{cases} f_i^{\min}, & f_i^* \le f_i^{\min} \\ \sqrt[3]{\dfrac{\sum_j a_{i,j}\lambda_{i,j}^*}{2\kappa\beta}}, & f_i^{\min} \le f_i^* \le f_i^{\max} \\ f_i^{\max}, & f_i^{\max} \le f_i^* \end{cases}$$

$$\ell_{i,j}^* = a_{i,j}\sqrt{\frac{qM(\beta p_{i,j} + \lambda_{i,j}^*)}{\gamma^* \log_2(1 + \frac{p_{i,j}h_{i,j}}{N_0})}}$$

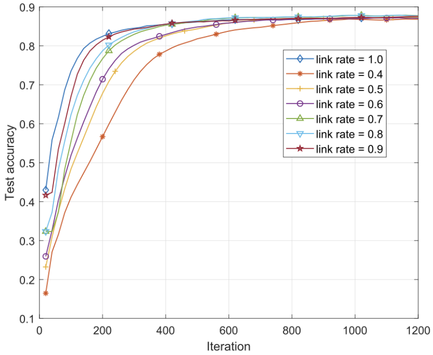
The $\lambda_{i,j}^*$ and γ^* are lagrange multipliers.

Outline

The Effect of the Link Number



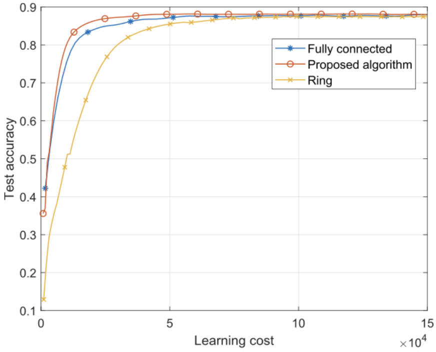
(a) Per-iteration cost



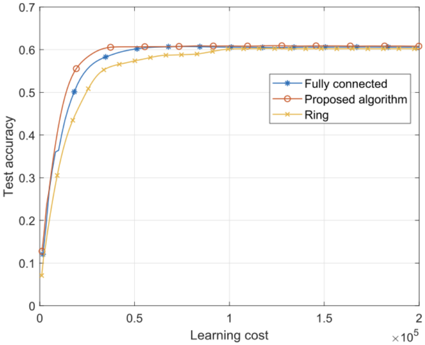
(b) Convergence rate

- The per-iteration learning cost increases with the link number.
- With more activated D2D links, the system requires less iterations to converge to the target accuracy due to a high network connectivity.

Performance Comparison of Link Selection Strategy



(c) ResNet, CIFAR10



(d) ResNet, CIFAR100

- The the proposed link selection mechanism outperforms other two schemes.
- Although the number of steps required for fully connected network convergence is the least, fully connected is not optimal, and the network needs to be sparse to reduce overhead.