ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 3,May2024,pp. 577－587
DOI:10. 16337/j. 1004-9037. 2024. 03. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于联邦分割学习与低秩适应的RoBERTa预训练模型微调方法

谢思静，文鼎柱

(上海科技大学信息科学与技术学院，上海 201210)

摘　要：微调后的大语言模型(Large language models，LLMs)在多任务中表现出色，但集中式训练存在用户隐私泄漏的风险。联邦学习(Federated learning，FL)通过本地训练避免了数据共享，但LLMs庞大的参数量对资源受限的设备和通信带宽构成挑战，导致在边缘网络中部署困难。结合分割学习(Split learning，SL)，联邦分割学习可以有效解决这一问题。基于模型深层权重的影响更为显著，以及对部分层的训练准确率略低于整体模型训练的发现，本文按照Transformer层对模型进行分割，同时引入低秩适应(Low-rank adaption，LoRA)进一步降低资源开销和提升安全性。因此，在设备端，仅对最后几层进行低秩适应和训练，然后上传至服务器进行聚合。为了降低开销并保证模型性能，本文提出了基于联邦分割学习与LoRA的RoBERTa预训练模型微调方法。通过联合优化边缘设备的计算频率和模型微调的秩，在资源受限的情况下最大化秩，提高模型的准确率。仿真结果显示，仅训练LLMs最后3层的情况下，在一定范围内(1～32)增加秩的取值可以提高模型的准确率。同时，增大模型每轮的容忍时延和设备的能量阈值可以进一步提升模型的准确率。

关键词：大语言模型；低秩适应；联邦学习；分割学习；联合优化

中图分类号：TP181　　**文献标志码：**A

Fine-Tuning Method for Pre-trained Model RoBERTa Based on Federated Split Learning and Low-Rank Adaptation

XIE Sijing，WEN Dingzhu

(School of Information Science and Technology，ShanghaiTech University，Shanghai 201210，China)

Abstract：Fine-tuned large language models (LLMs) perform exceptionally well in various tasks，but centralized training poses user privacy leakage risks. Federated learning (FL) mitigates data sharing issues through local training，yet the large parameter size of LLMs challenges resource-constrained devices and communication bandwidth，making deployment in edge networks difficult. Considering split learning (SL)，federated split learning can effectively address these issues. Given the more pronounced influence of deep-layer model weights and the discovery that training certain layers yields slightly lower accuracy compared to training the entire model，we opt to split the model based on Transformer layers. Additionally，utilizing low-rank adaption (LoRA) can further reduce resource overhead and enhance security. Therefore，at each device，we only perform LoRA and training on the final few layers. These adapted layers are then uploaded to the server for aggregation. From the perspective of cost reduction and

收稿日期：2024-03-30；**修订日期：**2024-04-27

ensuring model performance, we propose a fine-tuning method for the pre-trained model RoBERTa based on federated split learning and LoRA. By jointly optimizing the computational frequency of edge devices and the rank of model fine-tuning, we maximize the rank to improve model accuracy under resource constraints. Simulation results indicate that only training the last three layers of the LLMs can improve model accuracy within a certain range (1—32) by increasing the rank. Additionally, increasing the per-round delay and the energy threshold of devices can further enhance model accuracy.

Key words: large language models (LLMs); low-rank adaption (LoRA); federated learning (FL); split learning (SL); joint optimization

引　言

预训练大语言模型(Large language models,LLMs)自主提取语言内在的规则、语法和语义,拥有超强的语言理解和生成能力[1-4],在自然语言理解[5]、文本生成[6]等各类自然语言处理(Natural language processing, NLP)任务中实现了突破[7]。然而,随着LLMs规模的不断扩大,参数量已突破了万亿级,由模型规模带来的性能提升出现边际递减效应[8]。重新开始完整训练这些模型需要大量的计算资源和时间成本[9],对大多数设备难以实现。因此,针对特定任务的模型微调(Fine-tuning, FT)已成为利用LLMs的主要方法[10-11]。微调只需要关注模型的特定部分,在特定数据集上进一步调整已训练过的LLMs[12]。虽然以集中式框架进行微调是首选,但传统公开的可用数据集无法满足训练LLMs的需求[13]。然而,将所有数据聚合到单个设备上会引起难以容忍的时延和数据泄漏的隐私问题,从而使集中训练变得具有挑战性。此外,以用户为中心的个性化LLMs需求显著增加。

为了安全地利用大量分布式和私有数据训练LLMs,联邦学习(Federated learning, FL)[14]作为一种分布式框架被引入。联邦学习充分利用边缘设备的计算能力和存储资源,使模型能够跨设备进行训练。每个设备都基于本地样本数据进行训练,训练完成后上传模型参数或梯度至服务器进行聚合,而无需共享原始数据。端边协同实现更快速的模型训练和更低延迟的响应,同时减轻云服务器的负担,提高系统的整体性能和效率。然而,通过无线联邦学习微调LLMs面临着以下挑战。(1)联邦学习需要服务器和设备之间频繁通信传递模型参数或梯度来更新全局模型。LLMs参数数量庞大,导致巨大通信开销。此外,通信成本随着通信轮数、参与设备数量和模型大小的增加而增加。(2)LLMs微调需要大量的内存和计算资源,部分边缘设备可能会受到存储资源、计算能力和能量等方面的限制,导致训练速度慢或失败。(3)参与联邦学习的设备通常在有限的带宽网络上进行通信,导致数据传输出现巨大的延迟,从而降低模型训练效率。

为了解决上述挑战,尤其在边缘设备资源受限的情况下,联邦分割学习(Federated split learning, FSL)框架[15]被提出。联邦分割学习将模型分割成多个部分,在边缘设备上仅针对部分模型进行训练,并只传输更新部分的参数。在联邦分割学习框架下微调LLMs可以有效降低边缘设备和服务器之间传递的参数数量,从而减少各设备的计算和通信开销,提高了效率和隐私安全。在传统神经网络领域,通过观察网络不同层次的激活模式,发现随着网络深度的增加,特征和模式变得更加明显和高级[16]。在LLMs领域,在固定总体专家数量的情况下,将更多的LoRA专家分配给Transformer模型的更高层可以进一步提升模型的性能[17]。这表明更靠近输出层的权重控制着网络中更高层次的特征提取和决策过程,其重要性更为突出。因此,对于分割后的网络,训练后几层更为重要。

参数高效微调(Parameter efficient fine tuning, PEFT)[18-21]已成为一种替代训练策略,它只更新小部分特定于任务的参数。PEFT根据微调的参数性质分为两类:(1)微调已有的参数,而不添加额外的模块。Zaken等[18]提出BitFit(Bias-term fine-tuning)方法,只更新模型中的偏置参数,从而有效地简化

了参数选择的问题;(2)基于模块化的微调,引入与模型参数并行的模块。Li等[19]提出了前缀调优策略,在模型输入前添加一个连续且任务特定的向量序列(即前缀),只更新优化特定任务的前缀。低秩适应(Low-rank adaptation, LoRA)技术由Hu等[20]提出。LoRA将预训练模型的权重矩阵分解为两个低秩矩阵($\boldsymbol{A}$ 和 $\boldsymbol{B}$)的乘积。在训练过程中,预训练模型的权重($\boldsymbol{W}_0$)被冻结保持不变,而可训练的低秩矩阵($\boldsymbol{A}$ 和 $\boldsymbol{B}$)会根据下游任务需求进行调整。现有的工作已证明这些方法可以在减少训练所需参数预算的同时保证任务性能。在联邦分割学习架构中引入参数高效微调,能进一步降低每轮中需要更新的参数数量,提高模型的训练效率和增强系统的安全性。

本文探索了联邦分割学习架构下的LoRA方法。首先对LLMs按照Transformer层进行分割,选取分割后模型的最后3层进行低秩化处理并部署在边缘设备上进行联邦训练。每轮训练中仅对选中的层进行训练、上传、更新,以取代对整个模型的更新。对分割后的模型进行低秩化处理后,当秩取值为1~128时,参数量急剧下降,仅为原来模型参数量(129, 960, 968)的0.13%~1.02%。降低通信和计算开销的同时,为了保证模型性能,本文提出了一种基于FSL与LoRA的RoBERTa预训练模型微调方法。通过联合优化边缘设备的计算频率和模型微调的秩,在网络资源受限的情况下最大化秩,提高模型的准确率。本文的主要贡献如下:

(1) 基于不同层权重对网络的影响不同,研究了在大模型各位置采取LoRA后网络的性能差异。本文分别选取了4种层数选择方法:靠近数据输入端的前3层;中间任取的3层连续网络;靠近网络输出端的最后3层和完整12层网络进行对比。仅对选中的层采取LoRA并参与训练,同时将其他预训练权重冻结为其初始预训练值。

(2) 对联邦训练网络和系统模型进行建模,在网络计算和通信资源受限的约束下,设计了训练矩阵秩的优化问题。若秩取值过小会导致收敛速度慢,通过研究找到秩的最佳取值范围,以确保模型具有最大的表征能力,从而性能最高。由于秩的整数特性,使得优化问题成为一个混合整数线性规划问题,即非凸问题。通过整数约束放缩解决此非凸问题,然后针对获得的非整数解进行分支处理,从而获得最优解。

(3) 为了研究秩的取值和时延能量约束对系统性能的影响,本文使用RoBERTa模型和AGNews数据集进行仿真。在秩取值范围为1~32的情况下,发现随着秩的增加,模型的准确率也随之提高。然而,当秩超过一定阈值后,模型的准确率开始下降。这是因为过大的秩会引入冗余的参数信息,导致模型过度拟合。随着每轮延迟和设备能量约束的增加,模型的准确率持续上升。

1 系统模型与问题建立

1.1 系统模型

本文考虑一个单天线的服务器,$I=3$ 个边缘设备的网络模型架构如图1所示,图中 I 为设备的总数,$l_i(i=1,2,\cdots,I)$ 为第 i 个设备本地训练的层数,$\boldsymbol{A}$、$\boldsymbol{B}$ 为初始化的低秩矩阵。边缘设备是指连接到网络边缘能进行计算的移动设备,如手机、传感器等。由于在数据产生的地方可以直接进行计算和数据处理,边缘设备能减少数据传输,可拥有高实时性和可靠性。大语言模型的后3层部署在设备端参加每轮训练,其余层进行冻结保持不变。在每一轮中,设备首先利用本地数据集对后3层模型进行训练,随后将更新的模型参数上传至服务器端。服务器端对接收到的模型参数进行聚合更新后,再将新的模型参数广播给所有设备。以上步骤循环执行,直至全局准确率达到预设阈值或通信轮数达到提前设定的界限。不失一般性,文中假设信道具有频率选择性,且在每轮过程中信道保持静止,但在不同轮次中动态改变[22]。系统所有可用带宽为 B,在每一轮中被正交分为 $I=3$ 个正交子带,每个子带分配给一个设备进行参数上传和下载。服务器可以视为协调器,通过使用有效的信道估计方法获得上行链路和下行链路的信道状态信息(Channel state information, CSI)。另外,本文假设每个设备本地已存有原始的预训练模型权重。

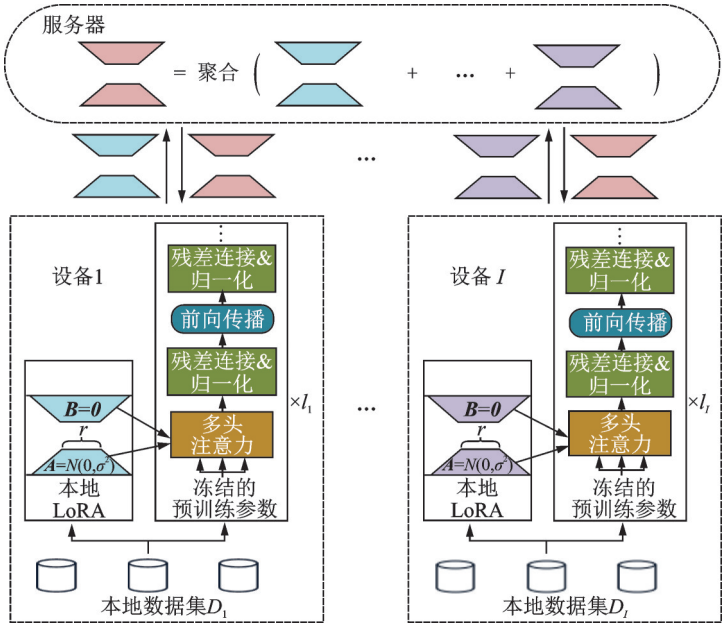

图1 基于FSL与LoRA的网络模型结构

Fig.1 Network model based on FSL and LoRA

1.2 时延与能源消耗模型

考虑任意一轮,即第t轮、第i个设备,各步骤产生的时间延迟和能量消耗分别建模为以下模型。其中,各参数代表的具体含义见表1。

表1 FSL与LoRA的参数名称及对应含义

Table 1 Parameters of FSL and LoRA and corresponding meanings

参数	含义	参数	含义
$C_{i,t}$	一个样本处理器更新权重的操作数	$M_{i,t}$	设备i上传的参数量
$\lvert\mathcal{D}_i\rvert$	设备i样本数	$s_{i,t}^{\mathrm{ul}}$	设备i上行链路数据传输率
$f_{i,t}/(\mathrm{cycle}\cdot\mathrm{s}^{-1})$	设备i运算频率	B_i	设备i的带宽
$\hat{A}$	1层中低秩矩阵的个数	$h_{i,t}^{\mathrm{ul}}$	设备i上行链路信道增益
l	本地更新的层数	$P_{i,t}^{\mathrm{ul}}$	设备i发送功率
r	低秩矩阵的秩	$s_{i,t}^{\mathrm{dl}}$	设备i下行链路数据传输率
Ω_i	设备i处理器计算性能的常数	$h_{i,t}^{\mathrm{dl}}$	设备i下行链路信道增益
$P_{i,t}^{\mathrm{dl}}$	服务器发送功率	ξ_i	设备i电路能量消耗

1.2.1 本地训练

第i个设备本地训练模型权重产生的时延可以建模为

$$T_{i,t}^{\mathrm{cmp}}=\frac{C_{i,t}|\mathcal{D}_i|}{f_{i,t}} \tag{1}$$

假设可训练的低秩矩阵$B\in\mathbf{R}^{d\times r}$,$A\in\mathbf{R}^{r\times k}$,完成一个元素的计算需要$r$次乘法和$r-1$次加法,则

$$C_{i,t}=(2r-1)\times d\times k\times\hat{A}\times l_i \tag{2}$$

本地计算的能量消耗[23]取决于$f_{i,t}$和$T_{i,t}^{\mathrm{cmp}}$,可表示为

$$E_{i,t}^{\mathrm{cmp}}=\Omega_i T_{i,t}^{\mathrm{cmp}}(f_{i,t})^3 \tag{3}$$

1.2.2 模型上传

第 i 个设备模型上传的时延可以表示为

$$T_{i,t}^{\mathrm{com,ul}}=\frac{M_{i,t}Q}{s_{i,t}^{\mathrm{ul}}} \tag{4}$$

式中 Q 为每个参数的量化比特数。$M_{i,t}$ 可以表示为

$$M_{i,t}=(r+k)\times d\times \hat{A}\times l_i \tag{5}$$

$s_{i,t}^{\mathrm{ul}}$ 可以表示为

$$s_{i,t}^{\mathrm{ul}}=B_i\log_2\left(1+\frac{h_{i,t}^{\mathrm{ul}}P_{i,t}^{\mathrm{ul}}}{N_0}\right) \tag{6}$$

式中 N_0 为信道噪声的方差。根据式(6),$P_{i,t}^{\mathrm{ul}}$ 表示为

$$P_{i,t}^{\mathrm{ul}}=\left(2^{\frac{s_{i,t}^{\mathrm{ul}}}{B_i}}-1\right)\frac{N_0}{h_{i,t}^{\mathrm{ul}}} \tag{7}$$

因此,此步骤的能量消耗为

$$E_{i,t}^{\mathrm{com,ul}}=P_{i,t}^{\mathrm{ul}}T_{i,t}^{\mathrm{com,ul}} \tag{8}$$

1.2.3 模型聚合和模型下发

全局模型聚合及更新发生在服务器,产生的时间延迟对所有设备保持相同,因此被忽略。此外,在模型聚合步骤中,边缘设备不消耗任何能量。

服务器将更新后的模型参数下发给第 i 个设备产生的时延可以建模为

$$T_{i,t}^{\mathrm{com,dl}}=\frac{M_{i,t}Q}{s_{i,t}^{\mathrm{dl}}} \tag{9}$$

式中 $s_{i,t}^{\mathrm{dl}}$ 表达式为

$$s_{i,t}^{\mathrm{dl}}=B_i\log_2\left(1+\frac{h_{i,t}^{\mathrm{dl}}P_{i,t}^{\mathrm{dl}}}{N_0}\right) \tag{10}$$

此步骤中产生的能量消耗为接收模型,被包含在电路能量消耗 ξ_i 中。

综上所述,第 t 轮、第 i 个设备产生的总延迟为

$$T_{i,t}=T_{i,t}^{\mathrm{cmp}}+T_{i,t}^{\mathrm{com,ul}}+T_{i,t}^{\mathrm{com,dl}} \tag{11}$$

第 t 轮中本地设备产生的总能量消耗为

$$E_{i,t}=E_{i,t}^{\mathrm{cmp}}+E_{i,t}^{\mathrm{com,ul}}+\xi_i \tag{12}$$

后文为表述简单,略去轮数标识 t。

1.3 问题建立

若随机某轮训练中低秩矩阵的秩过低,训练的参数过少,模型表征能力就会较差,从而导致训练收敛速度变慢[20,24,25]。为提升训练的收敛速度,本文设计了最大化低秩矩阵秩的目标函数,用于最大化模型的表征能力,即

$$\max_{\{f_i,r\}}\ r \tag{13}$$

由于网络资源有限,如:本地设备有限的计算能力、能量等,存在以下约束条件。

(1) 时延约束

每一轮的总时延应小于给定的最大值,根据式(11)可得

$$T_i^{\mathrm{cmp}} + T_i^{\mathrm{com,ul}} + T_i^{\mathrm{com,dl}} \leqslant T_0 \tag{14}$$

将式(1, 4, 9)代入,可得

$$\mathscr{C}_1: \frac{C_i|\mathcal{D}_i|}{f_i} + \frac{M_i Q}{s_i^{\mathrm{ul}}} + \frac{M_i Q}{s_i^{\mathrm{dl}}} \leqslant T_0 \tag{15}$$

(2) 能量约束

本地设备的能量有限,每一轮消耗的能量不能超过设备给定的阈值,根据式(12)可得

$$E_i^{\mathrm{cmp}} + E_i^{\mathrm{com,ul}} + \xi_i \leqslant E_{i,0} \tag{16}$$

将式(3,8)代入,可得

$$\mathscr{C}_2: \Omega_i T_i^{\mathrm{cmp}} (f_i)^3 + \frac{M_i Q P_i^{\mathrm{ul}}}{s_i^{\mathrm{ul}}} + \xi_i \leqslant E_{i,0} \tag{17}$$

(3) 计算能力约束

本地设备计算能力有限,即

$$\mathscr{C}_3: 0 \leqslant f_i \leqslant f_{i,\max} \tag{18}$$

(4) 秩约束

由于秩的定义,取值只能取正整数,即

$$\mathscr{C}_4: r \in \mathbf{Z}^+ \tag{19}$$

基于上述限制条件,优化问题表示为

$$\mathrm{P}_1: \begin{cases} \max\limits_{\{f_i\}, r} \ r \\ \text{subject to } \ \mathscr{C}_1 \sim \mathscr{C}_4 \end{cases} \tag{20}$$

2 算法设计

由于秩的取值只能为正整数,因此优化问题为混合整数线性规划问题(Mixed-integer linear programming, MILP),是非凸问题。首先将整数规划问题转化为相应的线性规划问题,将所有变量视为连续变量,求解得到线性规划的最优解。若线性规划的最优解是整数,则整数规划问题已找到最优解。若线性规划的最优解不是整数,则选取秩(在线性规划解中取非整数值的变量)作为分支变量,拆分为两个子问题,两个限制分别为秩的上取整和下取整。对每个子问题重复求解的过程,直到找到最优整数解。因此求解问题转化为

$$\mathrm{P}_2: \begin{cases} \max\limits_{\{f_i\}, r} \ r \\ \text{subject to } \ \mathscr{C}_1 \sim \mathscr{C}_3 \end{cases} \tag{21}$$

对于优化问题P_2,给定训练矩阵的秩r,各设备的计算频率$\{f_i\}$可通过约束条件$\mathscr{C}_1 \sim \mathscr{C}_3$求解得到。因此,本文采用了二分法求解来降低算法复杂度。初始训练矩阵的秩分别选为秩能取的最小和最大值,即为r^{lef}、r^{rig},分别表示为

$$r^{\mathrm{lef}} = 1 \tag{22}$$

$$r^{\mathrm{rig}} = \max\left\{ T_0 - \frac{C_i|\mathcal{D}_i|}{f_{i,\max}} - \frac{M_i Q}{s_i^{\mathrm{ul}}} - \frac{M_i Q}{s_i^{\mathrm{dl}}} \right\} \tag{23}$$

每次根据中间点的求解情况进行新的赋值,所有设备求解完后,选取所有设备中秩的最小值为整个网络的训练矩阵的秩,最后根据r计算$\{f_i\}$。具体过程如算法1所示。值得注意的是,每个设备都可以同时在本地利用二分法求解P_2,即该算法可以在分布式系统上并行处理。

算法 1 P_2的二分法求解

输入: $\{s_i^{\mathrm{dl}}\}$, $\{s_i^{\mathrm{ul}}\}$, T_0, 和 $\{E_{i,0}\}$

(1) for each device $i \in I$ do

(2) 初始化 $r_i^{\mathrm{lef}} \leftarrow r^{\mathrm{lef}}$, $r_i^{\mathrm{rig}} \leftarrow r^{\mathrm{rig}}$, $r_i^{\mathrm{mid}} \leftarrow \dfrac{r_i^{\mathrm{lef}} + r_i^{\mathrm{rig}}}{2}$

(3) while $r_i^{\mathrm{rig}} - r_i^{\mathrm{lef}} > 1$ do

(4) if $r = r_i^{\mathrm{mid}}$ has feasible solutions then

(5) $r_i^{\mathrm{lef}} \leftarrow r_i^{\mathrm{mid}}$

(6) $r_i^{\mathrm{mid}} \leftarrow \dfrac{r_i^{\mathrm{lef}} + r_i^{\mathrm{rig}}}{2}$

(7) else

(8) $r_i^{\mathrm{rig}} \leftarrow r_i^{\mathrm{mid}}$

(9) $r_i^{\mathrm{mid}} \leftarrow \dfrac{r_i^{\mathrm{lef}} + r_i^{\mathrm{rig}}}{2}$

(10) end

(11) $r_i^* \leftarrow r_i^{\mathrm{rig}}$

(12) end

(13) end

(14) $r \leftarrow \min\{r_i^*\}$

(15) 根据 $\mathscr{C}_1 \sim \mathscr{C}_3$,计算$\{f_i\}$

输出: r, $\{f_i\}$

3 实验结果与分析

3.1 实验环境

本文所有实验均在单台服务器,Linux系统下完成,服务器配置包含4块NVIDIA GeForce RTX 3090,每块显存为24 GB。网络结构基于Pytorch构建,编程语言为Python 3.8。本文仿真场景为1个单天线服务器和3个单天线设备的联邦分割学习系统。边缘设备和服务器之间的无线信道考虑大尺度衰落,采用瑞利衰落分布,瑞利衰落系数为10^{-3}。实验中将训练样本等分成3份并分别分配给每个设备,即每个设备含有40 000个新闻数据,其中每类数据10 000个。因此,每个设备分别具有独立同分布(Independent and identically distributed, IID)的数据集。其他相关仿真参数如表2所示,表中Unif表示均匀分布。

表 2 资源分配的仿真参数

Table 2 Simulation parameters of resource allocation

参数	数值
设备计算频率Ω_i/(cycle•s^{-1})	Unif(0.2,1)×10^{-25}
发射功率P/W	Unif(0.2,1)×0.1
量化比特数Q	256
带宽B_i/MHz	1
学习率	5e−5
本地更新轮数	1

3.2 模型和数据集

本文基于由Facebook AI Research开发的RoBERTa模型[26],是在Google的BERT模型基础上进行了改进。RoBERTa在预训练阶段舍弃了文本相似度任务,并引入了动态遮蔽策略,以增强模型对语言理解的能力。相比BERT,RoBERTa使用了规模更大的训练数据集,并且训练时间也更长,这虽然有助于模型捕捉更加丰富的语言特征,但也显著增加了训

练的计算开销。RoBERTa模型包含12层的Transformer编码器,每层后紧随的全连接层拥有768维的隐藏层,这一维度决定了模型在解析输入数据时能够捕捉到的表征复杂程度。

仿真中使用了AGNews数据集[27]。该数据集汇集了大量从新闻网站获取的新闻文章,现已被广泛应用于各种NLP任务,包括文本分类和情感分析等。AGNews数据集分为训练集和测试集,其中训练集包含约40万篇文章,总计12万个训练样本;测试集则含有约1.5万篇文章,共计7 600个测试样本。新闻内容涵盖4个主要类别,即世界新闻、体育、财经和科技,每个类别分别包含30 000个训练样本及1 900个测试样本,这为多种自然语言处理任务提供了强有力的基准数据支持。仿真实验中的测试准确率采用分类准确率来衡量,即分类正确的数据样本除以测试集的总数。

3.3 实验结果分析

3.3.1 低秩矩阵位置与准确率的关系

本节首先对比微调前后的测试准确率,完全未经过微调的模型准确率只有24.8%,低于秩取任意值情况下的微调,证明了微调的重要性。其次为了验证引言中低秩矩阵位置与准确率的关系,本节选取了4种大模型中层数位置结合低秩适应技术并进行对比,分别是靠近数据输入端的前3层,中间任取的3层连续网络,靠近网络输出端的最后3层和整个12层网络。结果如图2所示。可以看出,更新全部层数的准确率最高,这是因为所有的梯度信息可以在整个网络中自由传递和反向传播,同时整个模型在训练过程中保持一致性,避免了不同层之间的不稳定性。但本地设备和网络资源有限,在给定更新层数的前提下,更新网络的最后3层性能最好,和引言的推断一致。这是由于神经网络逐层传递信息的结构,更接近输出层(回归头)的权重承载着更多的信息,对最终输出结果有直接影响,从而更加重要。另外,只对部分层进行更新的准确率略低于所有层参与更新的准确率。

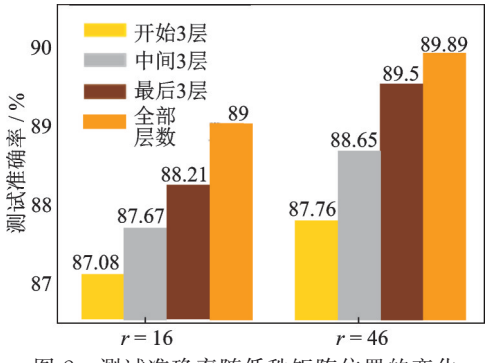

图2 测试准确率随低秩矩阵位置的变化

Fig.2 Variation of test accuracy with the location of low-rank matrix

3.3.2 准确率与秩的关系

本节研究了准确率与秩的关系,从图3中可以看出,在秩从1增长到32的过程中,准确率随之增长。然而,当秩进一步增加时,准确率不再提高,反而下降,表现出过拟合的现象。这表明适度增加训练矩阵的秩可以增强模型的表征能力和准确率。然而,过大的秩并不会引入更多有效信息,而会导致信息冗余,对准确率的改善作用有限。因此,合理增加秩可以提高模型性能,但过度增加秩可能会产生负面影响,降低准确率并引发过拟合问题。这和Hu等[20]提出的观点一致,并验证了在网络资源受限的情况下,最大化训练矩阵秩的合理性。

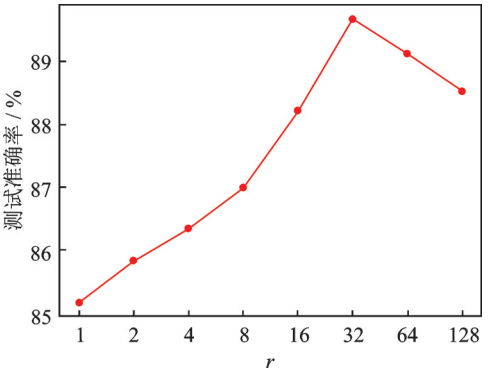

图3 测试准确率随秩的变化

Fig.3 Variation of test accuracy with rank

3.3.3 准确率与网络资源的关系

为了证明所提出的优化问题和优化算法的合理性,将提出的优化算法与随机设备计算频率f_i进行对

比,即在每轮中随机选取f_i,再对矩阵的秩进行优化。实验结果如图4和图5所示。

图4是每轮时延不同的情况下,测试准确率的变化情况。所提出的优化算法准确率一直高于随机设备计算频率的算法,因为联合优化计算频率与秩可以更高效地利用所有资源,使秩的取值最大。另外,从图4中可以看出,随着每轮可容忍最大时延的增加,准确率与之上升。这是由于随着时延约束的放松,其余参数不变的情况下,更多的模型参数能够参与训练,即秩的取值可以更大,泛化性能更强,从而提升系统的准确率。

图5是设备能量约束不同的情况下,测试准确率的变化情况。所提出的优化算法准确率一直高于随机设备计算频率的算法,证明了算法的有效性。和时延约束相同,随着设备能量阈值的提升,系统的准确率也随之增加,原因和上一种时延情况相同,故此处省略。

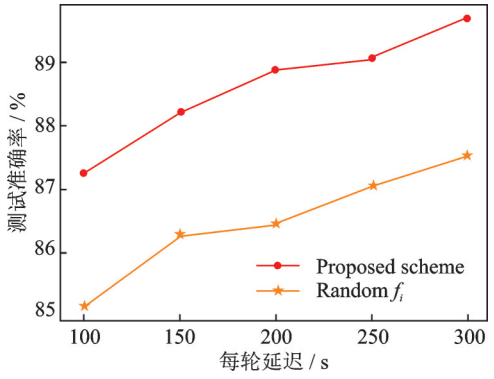

图4 测试准确率每轮延迟的变化

Fig.4 Variation of test accuracy with the per-round latency

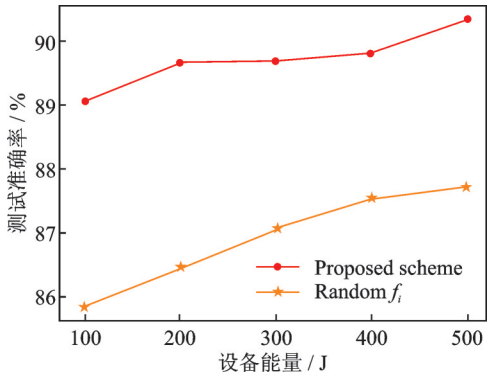

图5 测试准确率随设备能量的变化

Fig.5 Variation of test accuracy with device energy

4 结束语

为了缓解大型语言模型中庞大参数量带来的计算和通信挑战,本文在联邦分割学习框架中采用了大语言模型中的低秩适应技术,冻结大部分层模型参数,实现了在网络资源有限的约束下最大化模型准确率的要求。本文提出了一个有效的联合优化算法,为边缘设备选取合适的计算频率,最大程度上保障所选中模型的部分能更多地参与训练,从而提高模型表征能力。分析仿真结果可以发现,本文提出的算法能够有效提高分类任务的准确性。为了适应各边缘设备之间的差异,未来可以考虑动态异构地调整低秩矩阵的秩和各边缘设备参与训练的层数。同时,改变边缘设备的数据分布和增加多个多天线的服务器,使场景更加接近实际生活中的应用。

参 考 文 献:

[1] HU Linmei, LIU Zeyi, ZHAO Ziwang, et al. A survey of knowledge enhanced pre-trained language models[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(4): 1413-1430.

[2] RAIAAN M A K, MUKTA M S H, FATEMA K, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges[J]. IEEE Access, 2024, 12: 26839-26874.

[3] FAN L, LI L, MA Z, et al. A bibliometric review of large language models research from 2017 to 2023[EB/OL]. (2023-04-09). https://arxiv.org/abs/2304.02020.

[4] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[EB/OL]. (2023-03-15). https://arxiv.org/abs/2303.18223.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding

[EB/OL]. (2018-10-10). https://arxiv.org/abs/1810.04805.

[6] ZENG A, LIU X, DU Z, et al. GLM-130B: An open bilingual pre-trained model[EB/OL]. (2022-10-01). https://arxiv.org/abs/2210.02414.

[7] 夏润泽, 李丕绩. ChatGPT 大模型技术发展与应用[J]. 数据采集与处理, 2023, 38(5): 1017-1034.
XIA Runze, LI Piji. Large language model ChatGPT: Evolution and application[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1017-1034.

[8] 罗锦钊, 孙玉龙, 钱增志, 等. 人工智能大模型综述及展望[J]. 无线电工程, 2023, 53(11): 2461-2472.
LUO Jinzhao, SUN Yulong, QIAN Zengzhi, et al. Overview and prospect of artificial intelligence large models[J]. Radio Engineering, 2023, 53(11): 2461-2472.

[9] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient large-scale language model training on GPU clusters using megatron-LM[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ,USA: IEEE, 2021: 1-15.

[10] DING N, QIN Y, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.

[11] 沈之杰, 郭武. 基于预训练与音素字节对编码的越南语识别[J]. 数据采集与处理, 2023, 38(1): 101-110.
SHEN Zhijie, GUO Wu. Vietnamese speech recognition based on pre-training and phone-based byte-pair encoding[J]. Journal of Data Acquisition and Processing, 2023, 38(1): 101-110.

[12] COLIN R. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020. DOI: 10.48550v arxiv.1910.10683.

[13] JIANG F, DONG L, TU S, et al. Personalized wireless federated learning for large language models[EB/OL]. (2024-04-01). https://arxiv.org/abs/2404.13238.

[14] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//Proceedings of Artificial Intelligence and Statistics. [S.l.]: PMLR, 2017: 1273-1282.

[15] THAPA C, ARACHCHIGE P C M, CAMTEPE S, et al. Splitfed: When federated learning meets split learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 8485-8493.

[16] YOSINSKI J, CLUNE J, NGUYEN A, et al. Understanding neural networks through deep visualization[EB/OL]. (2015-06-18). https://arxiv.org/abs/1506.06579.

[17] GAO C, CHEN K, RAO J, et al. Higher layers need more LORA experts[EB/OL]. (2024-02-20). https://arxiv.org/abs/2402.08562.

[18] ZAKEN E B, GOLDBERG Y, RAVFOGEL S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 1-9.

[19] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 4582-4597.

[20] HU E J, SHEN Y, WALLIS P, et al. LORA: Low-rank adaptation of large language models[EB/OL]. (2021-06-21). https://arxiv.org/abs/2106.09685.

[21] BABAKNIYA S, ELKORDY A R, EZZELDIN Y H, et al. SLoRA: Federated parameter efficient fine-tuning of language models[C]//Proceedings of International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023. Cambridge, MA, USA: OpenReview, 2023.

[22] WEN D, BENNIS M, HUANG K. Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(12): 8272-8286.

[23] YOU C, HUANG K, CHAE H, et al. Energy-efficient resource allocation for mobile-edge computation offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3): 1397-1411.

[24] CHO Y J, LIU L, XU Z, et al. Heterogeneous LORA for federated fine-tuning of on-device foundation models[C]// Proceedings of International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023. Cambridge, MA,USA: OpenReview, 2023.

[25] ZHANG Q, CHEN M, BUKHARIN A, et al. Adaptive budget allocation for parameter-efficient fine-tuning[C]//Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda: DBLP, 2023.

[26] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-01). https: //arxiv.org/abs/1907.11692.

[27] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[J]. Advances in Neural Information Processing Systems, 2015. DOI: 10.48850/arxiv.1509.01626.

作者简介:

谢思静(2000-),女,硕士研究生,研究方向:无线联邦学习,E-mail: Xiesj2023@shanghaitech.edu.cn。

文鼎柱(1992-),**通信作者**,男,助理教授,博士生导师,研究方向:边缘人工智能、通信-感知-计算融合,E-mail: wendzh@shanghaitech.edu.cn。

(编辑:张黄群)