ICML
International Conference
On Machine Learning

Federated Sketching LoRA: A Flexible Framework for Heterogeneous Collaborative Fine-Tuning of LLMs

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan

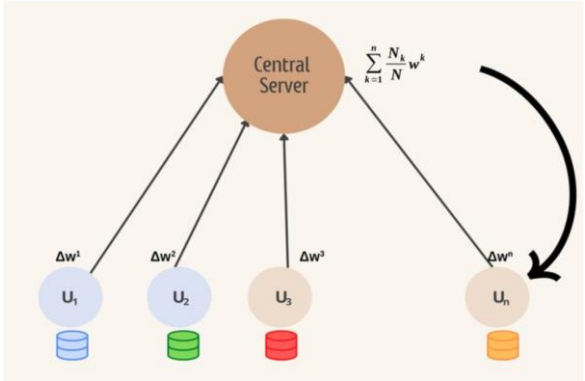
Seyyedali Hosseinalipour, Christopher G. Brinton

Motivation I: On-Device LLM

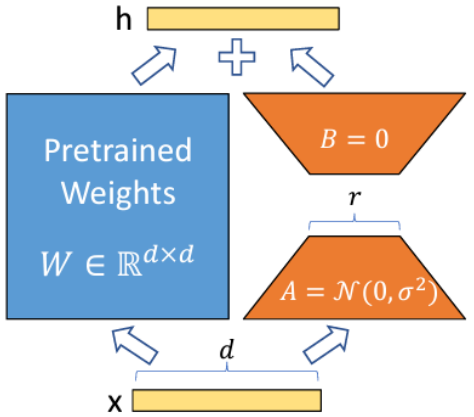
- Challenges of Cloud-based LLMs
  - **High Latency**: Real-time applications suffer from delays due to network dependency
  - **Privacy Concerns**: User data is transmitted to remote servers, risking data leakage
  - **Underused Device Power**: The potential of device is underexplored
- Advantages of on-device LLMs (Qualcomm & Apple AI):
  - Offline Capability
  - Enhancing Privacy
  - Better Personalization

Motivation II: Distributed On-Device LLM

- Fine-tuning is required for some specific downstream task
  - **Domain Adaptation**: aligns the model with domain-specific language (e.g., healthcare, legal)
  - **Task Specialization**: tailors the model for different tasks translation, summarization, and etc..
- Limited data:
  - ➢mobile device own limited data, which is enough for fine-tuning
  - ➢solution: collaborative fine-tuning



- Model transmission will induce communication burden
  - ➢LoRA is a solution

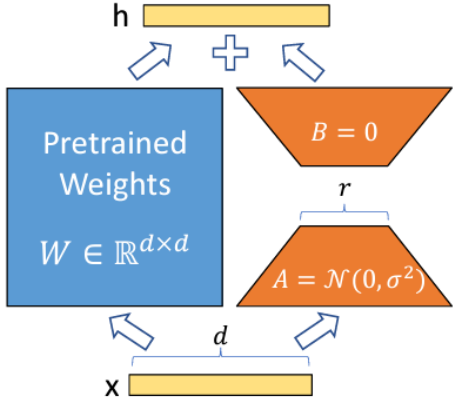


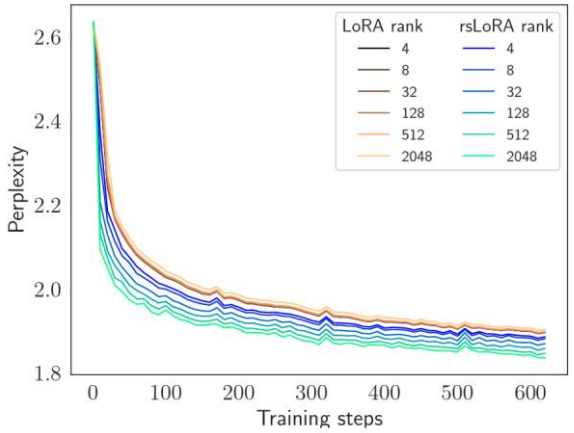
Motivation III: Resource heterogeneity

- LoRA's Limitaion
  - Uniform rank is necessary for *aggregation*,
  - The communication and computation resource is *heterogeneous* across edge devices
  - Large rank could still induce *communication and computation* burden

$$\min_{\mathbf{B},\mathbf{A}} f(\mathbf{B},\mathbf{A}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{B},\mathbf{A})$$

$$f(\mathbf{B},\mathbf{A}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\ell(\mathbf{W}_0+\mathbf{BA},\xi)\right]$$

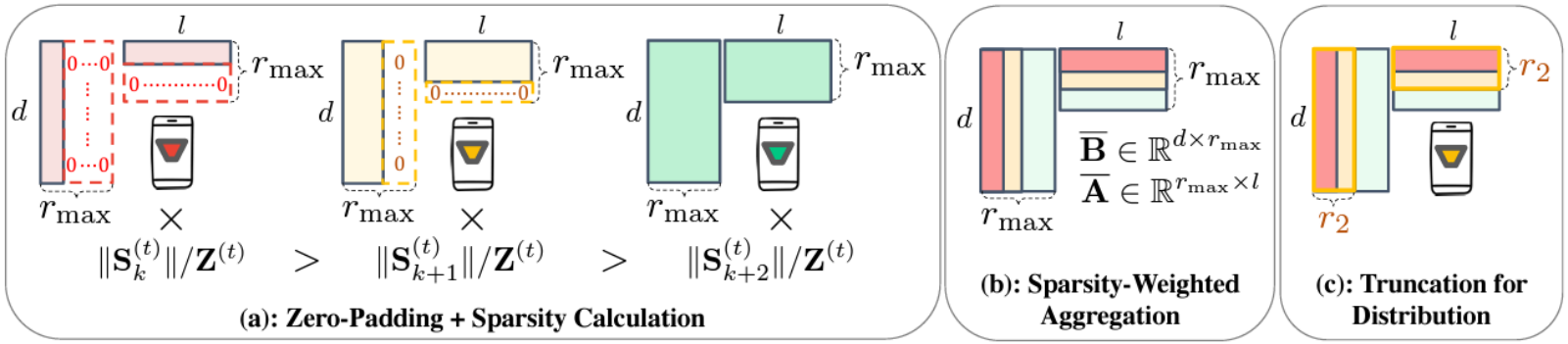




- How to enable heterogeneous LoRA with convergence guarantee?

Existing Solution

- How to enable heterogeneous LoRA with convergence guarantee?
  - Existing solution



  - Shortcoming: heuristic, without convergence guarantee

Sketching LoRA (a theoretical form of sub-LoRA partitioning)

- New formulation

$$\min_{\mathbf{B},\mathbf{A}} f(\mathbf{B},\mathbf{A}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{B},\mathbf{A})$$
$$f(\mathbf{B},\mathbf{A}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\ell(\mathbf{W}_0+\mathbf{BA},\xi)\right]$$

$$\min_{\mathbf{B},\mathbf{A}} f^{\mathcal{S}}(\mathbf{B},\mathbf{A}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_i^{\mathcal{S}}(\mathbf{B},\mathbf{A})$$
$$f_i^{\mathcal{S}}(\mathbf{B},\mathbf{A}) \coloneqq \mathbb{E}_{\mathbf{S}\sim\mathcal{S}_i;\xi\sim\mathcal{D}_i}\left[\ell(\mathbf{W}_0+\mathbf{BSA},\xi)\right]$$

- ***S*** is sketching matrix (sparse and diagonal), which is sampled from

$$\mathcal{S}(r,k)=\left\{\mathbf{S}\,|\,\mathbf{S}=\frac{r}{k}\sum_{j\in\mathcal{I}}\mathbf{e}_j\mathbf{e}_j^\top,\ \mathcal{I}\subseteq\{1,\dots,r\},\ |\mathcal{I}|=k\right\}$$

- Sparse gradient

$$\nabla_{\mathbf{B}}\tilde{\ell}(\mathbf{B},\mathbf{A},\xi)=\nabla\ell(\mathbf{W_0}+\mathbf{BSA},\xi)\mathbf{A}^\top\mathbf{S}^\top$$
$$\nabla_{\mathbf{A}}\tilde{\ell}(\mathbf{B},\mathbf{A},\xi)=\mathbf{S}^\top\mathbf{B}^\top\nabla\ell(\mathbf{W_0}+\mathbf{BSA},\xi)$$

Example: ***r = 5, k = 2***

$$S=\begin{bmatrix}0&0&0&0&0\\0&\frac{5}{2}&0&0&0\\0&0&0&0&0\\0&0&0&\frac{5}{2}&0\\0&0&0&0&0\end{bmatrix}$$

Sketching LoRA

- New formulation

$$\min_{\mathbf{B},\mathbf{A}} f(\mathbf{B},\mathbf{A}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{B},\mathbf{A})$$
$$f(\mathbf{B},\mathbf{A}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\ell(\mathbf{W}_0+\mathbf{BA},\xi)\right]$$

$$\min_{\mathbf{B},\mathbf{A}} f^{\mathcal{S}}(\mathbf{B},\mathbf{A}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_i^{\mathcal{S}}(\mathbf{B},\mathbf{A})$$
$$f_i^{\mathcal{S}}(\mathbf{B},\mathbf{A}) \coloneqq \mathbb{E}_{\mathbf{S}\sim\mathcal{S}_i;\xi\sim\mathcal{D}_i}\left[\ell(\mathbf{W}_0+\mathbf{BSA},\xi)\right]$$

- Sparse gradient

$$\nabla_{\mathbf{B}}\tilde{\ell}(\mathbf{B},\mathbf{A},\xi) = \nabla\ell(\mathbf{W_0}+\mathbf{BSA},\xi)\mathbf{A}^\top\mathbf{S}^\top$$
$$\nabla_{\mathbf{A}}\tilde{\ell}(\mathbf{B},\mathbf{A},\xi) = \mathbf{S}^\top\mathbf{B}^\top\nabla\ell(\mathbf{W_0}+\mathbf{BSA},\xi)$$

- Updating submatrix each iteration at edge devices!!

$$\left[\mathbf{B}_i^{t,h+1};\mathbf{A}_i^{t,h+1}\right] = \left[\mathbf{B}_i^{t,h};\mathbf{A}_i^{t,h}\right] - \gamma\left[\Delta\mathbf{B}_i^{t,h}(\mathbf{S}_i^t)^\top;(\mathbf{S}_i^t)^\top\Delta\mathbf{A}_i^{t,h}\right]$$

$$\left[\Delta\mathbf{B}_i^{t,h};\Delta\mathbf{A}_i^{t,h}\right] = \left[\nabla\ell(\mathbf{W}_0+\mathbf{B}_i^{t,h}\mathbf{S}_i^t\mathbf{A}_i^{t,h},\xi_i^{t,h})(\mathbf{A}_i^{t,h})^\top;(\mathbf{B}_i^{t,h})^\top\nabla\ell(\mathbf{W}_0+\mathbf{B}_i^{t,h}\mathbf{S}_i^t\mathbf{A}_i^{t,h},\xi_i^{t,h})\right]$$

Convergence analysis

- Results

Theorem 4.4. *Suppose that Assumptions 4.1-4.3 hold, then there exists a learning rate* $\gamma \leq \min\{\frac{N}{24\rho(c_h+1)H\bar{L}}, \frac{1}{8\sqrt{\widetilde{L}L(\rho+1)(c_h+1)H}}, \frac{1}{H}\}$ *such that the iterates* $\{\mathbf{X}^t\}$ *generated by FSLoRA satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla_{\mathbf{x}} f^{\mathcal{S}}(\mathbf{X}^t)\right\|^2 \leq 8\frac{\sqrt{\bar{L}\mathcal{F}_0\sigma_\rho^2}}{\sqrt{NTH}} + 10(\widetilde{L}L)^{\frac{1}{3}}\left(\frac{\mathcal{F}_0\sigma_\rho}{T}\right)^{\frac{2}{3}} + \frac{4\mathcal{F}_0}{T},$$

where $\sigma_\rho^2 = \sigma^2 + 3(\rho+1)\sigma_h^2$, $\bar{L} = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{r}{k_i}\right)L$, $\widetilde{L} = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{r^2}{k_i^2}\right)L$ *and* $\mathcal{F}_0 = f^{\mathcal{S}}(\mathbf{X}^0) - f^*$ *with* f^* *denoting the lower bound of* $f^{\mathcal{S}}(\mathbf{X})$.

- Converge to a stationary point
- Achieve speedup in the number of local iterations
- Recover to vanilla federated LoRA when $\boldsymbol{k} = \boldsymbol{r}$

Without non-diminishing bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left|\nabla_X f^S(X^t)\right|^2\right] \leq \mathcal{O}\left(\sqrt{\frac{\bar{L}}{NTH}}\right)$$

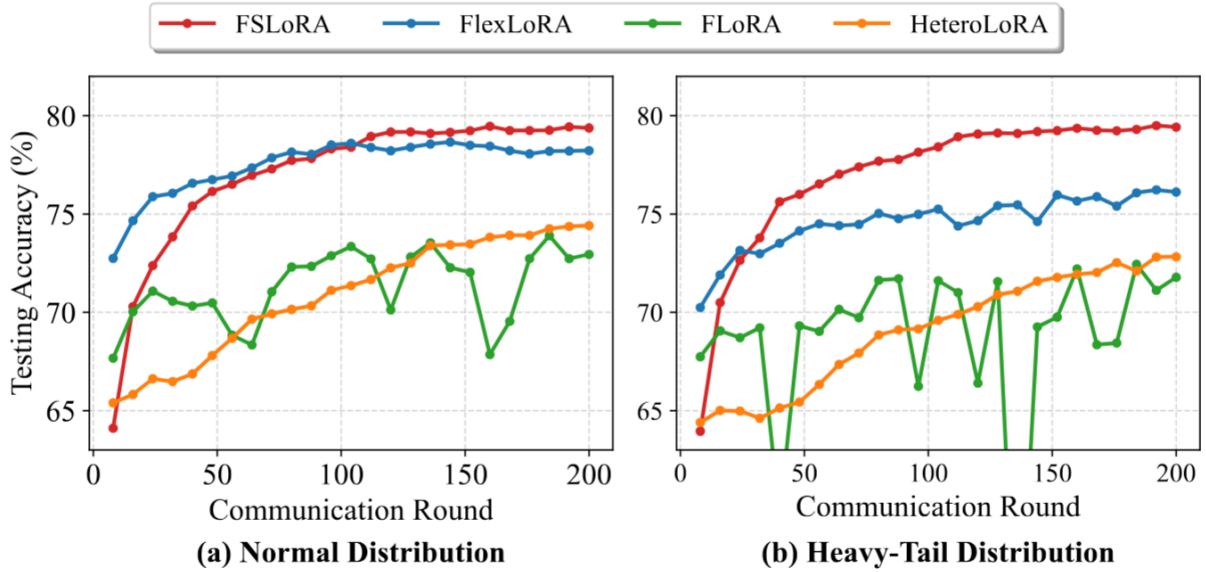
Experiment results: comparison with baselines

GLUE benchmark (RoBERTa model)

Method	GPU hours	QNLI	MRPC	CoLA	MNLI	RTE	SST-2	QQP	Avg.
HeteroLoRA	10.7h	87.5 ±0.5	84.4 ±0.9	75.3 ±1.2	66.3 ±0.8	69.0 ±1.7	89.5 ±0.0	85.3 ±0.1	79.6
FlexLoRA	12.6h	88.5 ±0.2	81.2 ±0.4	77.5 ±1.2	63.0 ±0.5	62.2 ±1.9	92.8 ±0.4	87.4 ±0.1	78.9
FLoRA	12.3h	87.2 ±0.3	78.1 ±0.7	77.4 ±1.7	74.6 ±0.5	54.4 ±2.1	93.4 ±0.1	87.1 ±0.3	78.9
FSLoRA	10.9h	88.0 ±0.3	87.3 ±0.2	82.2 ±0.5	76.4 ±0.2	69.8 ±1.2	93.5 ±0.1	85.8 ±0.1	83.3

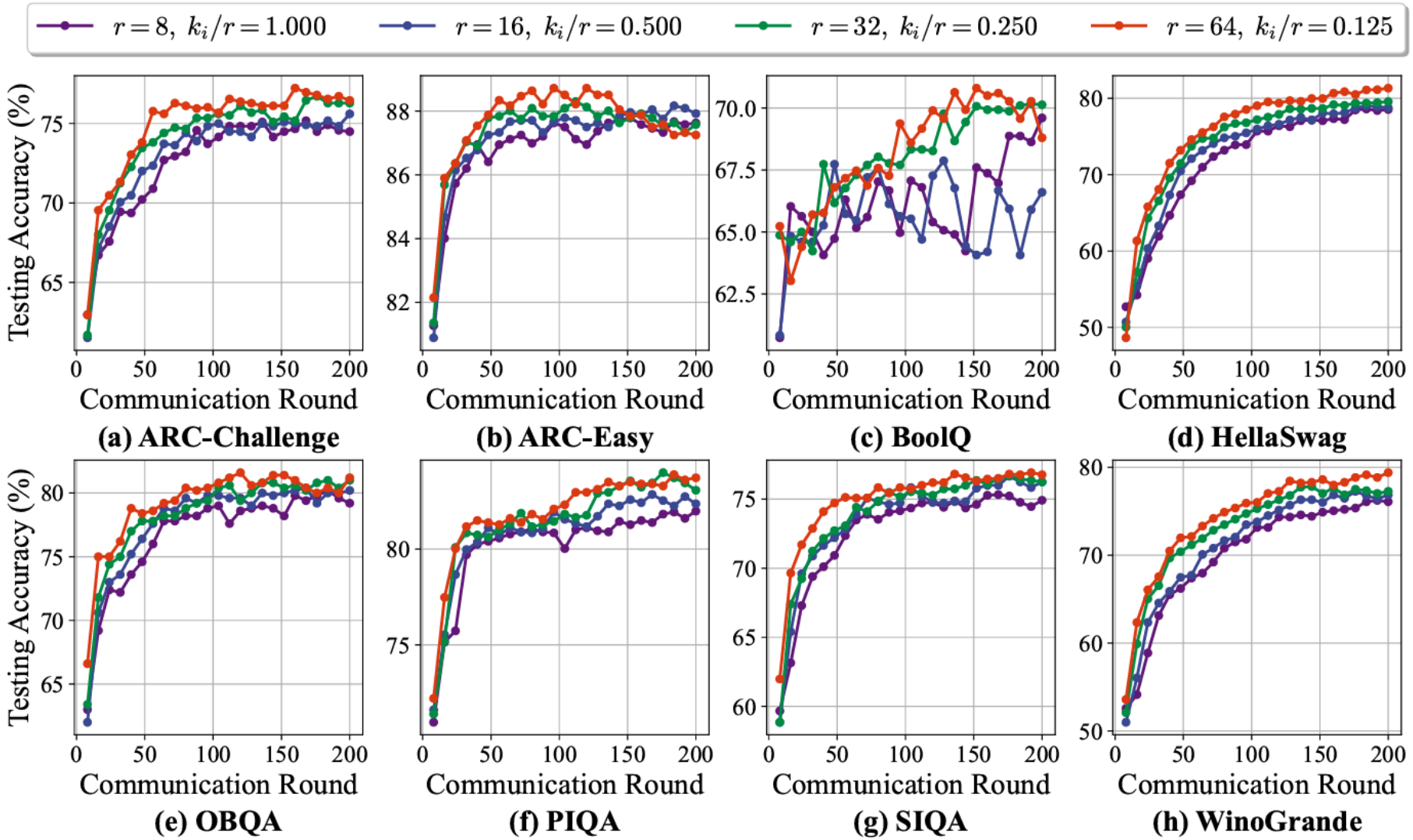
Commonsense reasoning benchmark (LLaMA-3.2-3B model)

Method	GPU hours	ARC-c	ARC-e	BoolQ	HellaSwag	OBQA	PIQA	SIQA	WinoGrande	Avg.
HeteroLoRA	43.7h	73.4 ±0.3	86.6 ±0.2	65.8 ±0.5	73.0 ±0.5	71.4 ±0.3	80.9 ±0.7	73.8 ±0.3	72.0 ±0.3	74.6
FlexLoRA	68.3h	74.2 ±0.3	86.7 ±0.6	68.6 ±0.8	79.4 ±0.7	75.8 ±0.4	81.0 ±0.3	75.9 ±0.4	77.9 ±0.3	77.4
FLoRA	49.8h	68.3 ±0.6	83.1 ±0.5	65.8 ±0.9	77.2 ±0.5	74.2 ±0.3	80.5 ±0.6	76.1 ±0.5	71.5 ±0.5	74.6
FSLoRA	44.3h	76.1 ±0.4	87.2 ±0.5	69.3 ±0.7	82.2 ±1.1	80.7 ±0.6	84.0 ±0.2	76.8 ±0.0	79.1 ±0.2	79.4



1. More stable convergence
2. Less GPU hour cost
3. Higher accuracy

Experiment results: superiority of sketching



Local ranks are the same, while the global ranks varies