DoFIT: Domain-aware Federated Instruction Tuning with Alleviated Catastrophic Forgetting

Binqian Xu[1], **Xiangbo Shu**[1,*], **Haiyang Mei**[2], **Zechen Bai**[2], **Basura Fernando**[3], **Mike Zheng Shou**[2], **and Jinhui Tang**[1]

[1]Nanjing University of Science and Technology [2]Show Lab, National University of Singapore
[3]Institute of High-Performance Computing, A*STAR

https://github.com/1xbq1/DoFIT

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

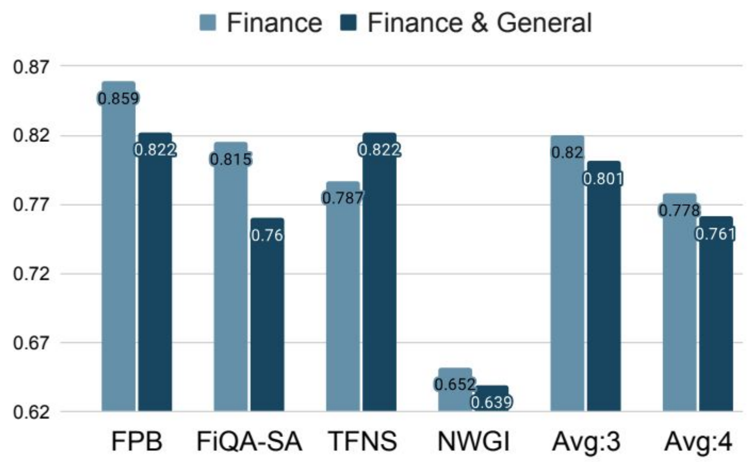
Motivation and Contribution

Motivation

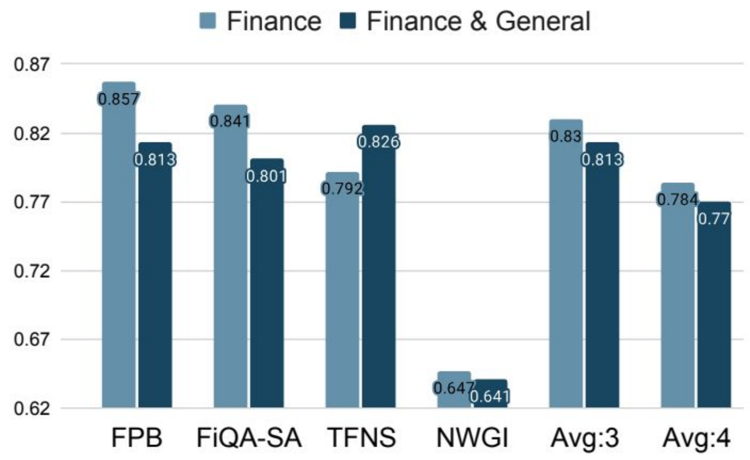
- Existing Federated Instruction Tuning (FIT) methods handle client-aware data heterogeneity, while ignoring domain-aware data heterogeneity.
- FIT: domain-information catastrophic forgetting problem.

Contribution

- A Domain-aware FIT baseline (DoFIT-base): finely aggregates overlapping important weights across domains to reduce interference.
- DoFIT with a new initialization strategy: incorporates inter-domain information into a less-conflicted parameter space to retain more domain information.

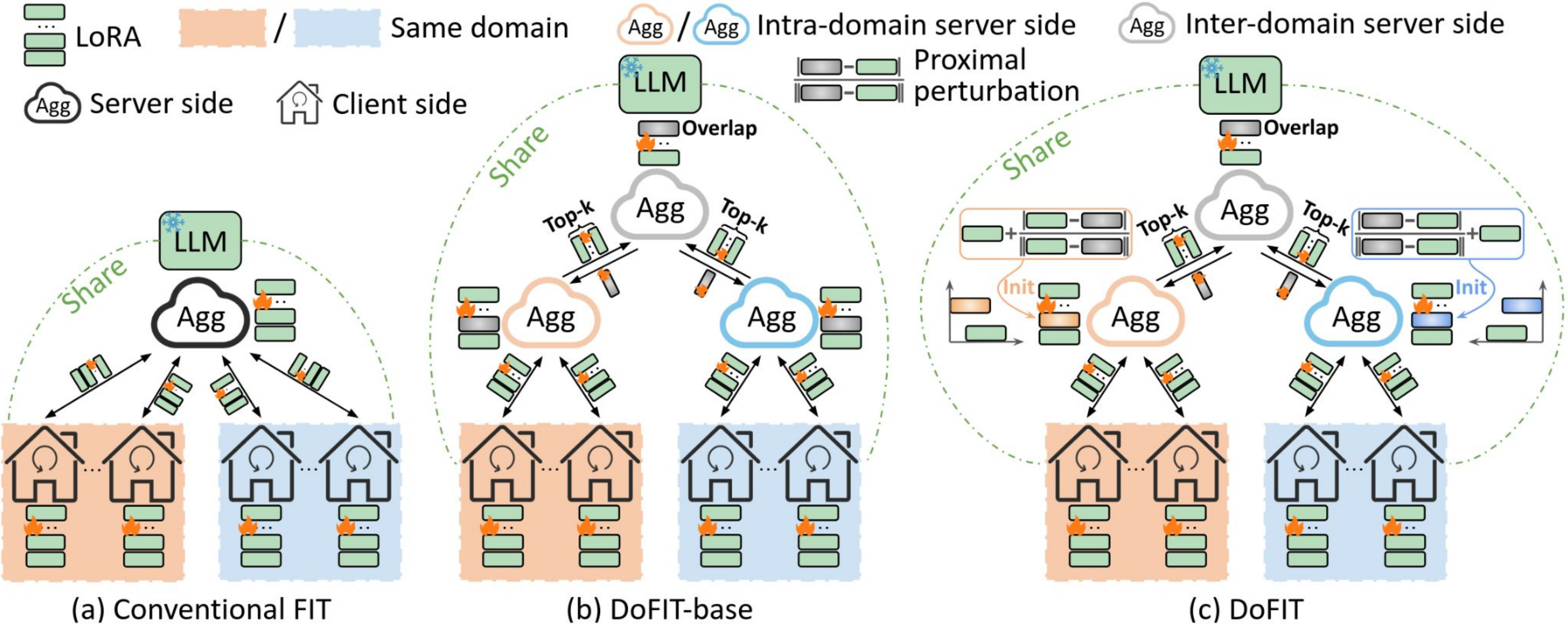


(a) Acc



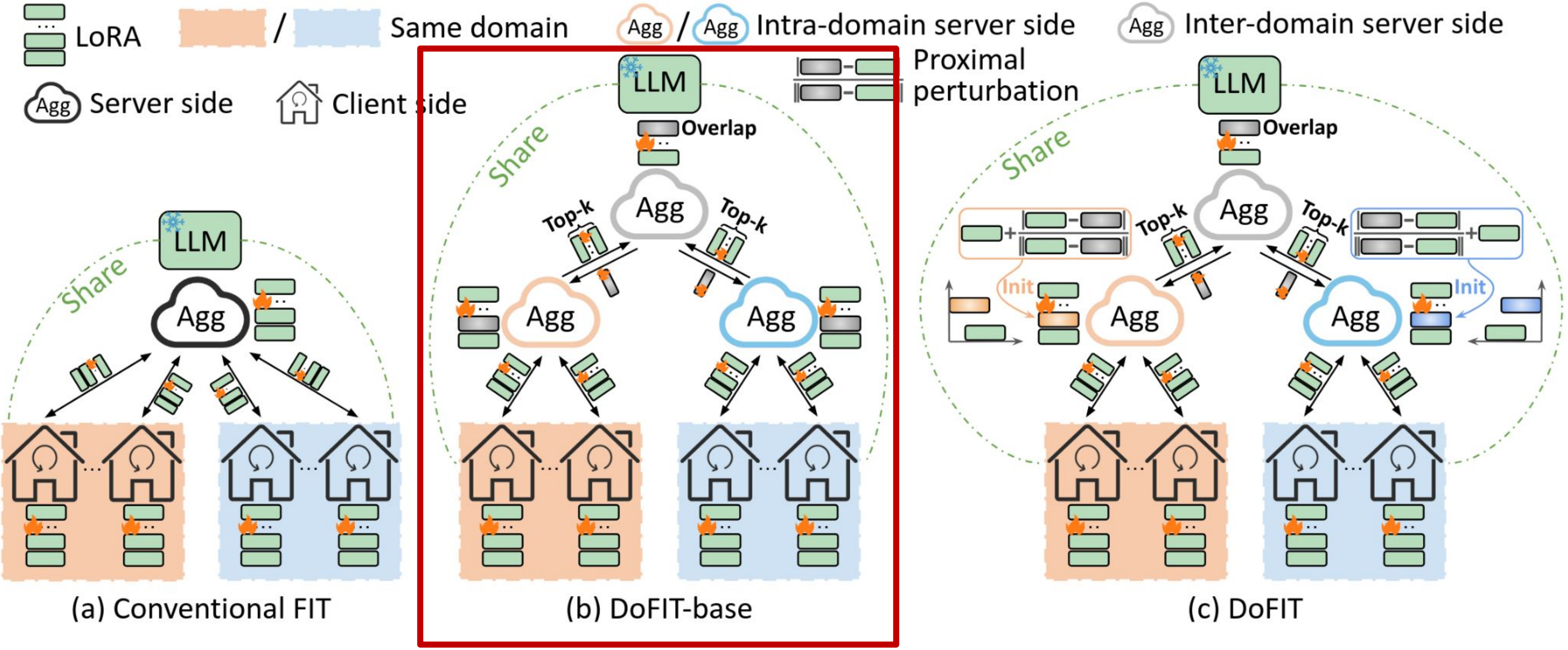
(b) F1

Overview



LoRA
/ Same domain
Agg / Agg Intra-domain server side
Agg Inter-domain server side
Agg Server side
Client side
Proximal perturbation
LLM
Share
Agg
Overlap
Top-k
Init
(a) Conventional FIT
(b) DoFIT-base
(c) DoFIT

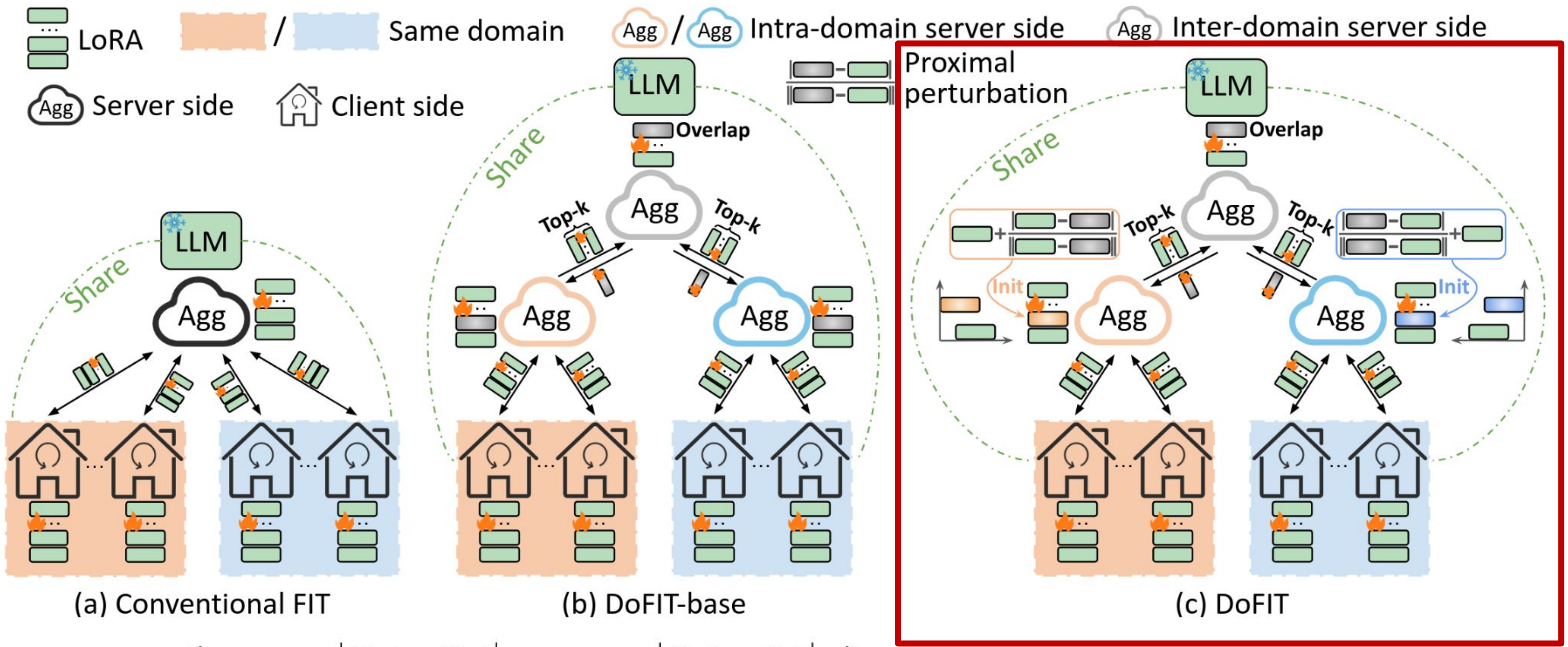
Overview



Compute the important score of each module for $\triangle \bar{W}_m^t$.
Upload top-k modules for $\triangle \bar{W}_m^t$ to the inter-domain server side.

important score: the squared norm of the module

Overview



$$\{\bar{B}_{m,l}^{t-1}\bar{A}_{m,l}^{t-1}\}_{\text{init}} = \left\{ (\bar{B}_{m,l}^{t-1} + \alpha \frac{\left|\widetilde{B}_{l}^{t-1} - \bar{B}_{m,l}^{t-1}\right|}{||\widetilde{B}_{l}^{t-1} - \bar{B}_{m,l}^{t-1}||_2})(\bar{A}_{m,l}^{t-1} + \alpha \frac{\left|\widetilde{A}_{l}^{t-1} - \bar{A}_{m,l}^{t-1}\right|}{||\widetilde{A}_{l}^{t-1} - \bar{A}_{m,l}^{t-1}||_2}) \right\}_{\in \Psi_{t-1}} \quad (10)$$

Experiments

Domain	Medthod	FPB		FiQA-SA		TFNS		NWGI		Avg:3		Avg:4	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
F	GPT-3.5	0.781	0.781	0.662	0.730	0.731	0.736	-	-	0.725	0.749	-	-
	GPT-4	0.834	0.833	0.545	0.630	0.813	0.808	-	-	0.731	0.757	-	-
	Local	0.770	0.760	0.655	0.719	0.742	0.747	0.629	0.624	0.722	0.742	0.699	0.713
	FIT_{32qv}	0.859	0.857	**0.815**	**0.841**	0.787	0.792	**0.652**	**0.647**	0.820	0.830	0.778	0.784
	FIT_{16qvd}	0.850	0.846	0.818	0.842	0.823	0.823	0.646	0.643	0.830	0.837	0.784	0.789
	FIT_{32d}	**0.860**	**0.857**	0.807	0.836	**0.824**	**0.825**	0.635	0.635	**0.830**	**0.839**	**0.782**	**0.788**
F&G	FIT_{32qv}	0.822	0.813	0.760	0.801	0.822	0.826	0.639	0.641	0.801↓	0.813↓	0.761↓	0.770↓
	Base_{top10}	0.859	0.855	0.778	0.815	0.810	0.811	0.637	0.638	0.816	0.827	0.771	0.780
	Base_{top15}	0.862	0.860	0.804	0.834	0.857	0.858	0.639	0.639	0.841↑	0.851↑	0.791↑	0.798↑
	Base_{top20}	0.859	0.855	0.775	0.815	0.866	0.864	0.632	0.634	0.833	0.845	0.783	0.792
	$\text{DoFIT}_{\alpha=0.5}$	**0.865**	**0.861**	0.815	0.842	0.864	0.864	**0.645**	**0.644**	0.848	0.856	0.797	0.803
	$\text{DoFIT}_{\alpha=1.0}$	0.861	0.858	**0.818**	**0.847**	**0.869**	**0.869**	0.641	0.640	**0.849**↑	**0.858**↑	**0.797**↑	**0.804**↑
	$\text{DoFIT}_{\alpha=1.5}$	0.859	0.855	0.815	0.845	0.822	0.825	0.642	0.641	0.832	0.842	0.785	0.792

Experiments

Domain	Method	MedQA	MedMCQA	Avg
M	Local	0.141	0.204	0.173
	FIT_{32qv}	0.167	**0.216**	**0.192**
	FIT_{16qvd}	**0.179**	0.200	0.190
	FIT_{32d}	0.158	0.199	0.179
M&G	FIT_{32qv}	0.174 $\uparrow$0.007	0.217 $\uparrow$0.001	0.196 $\uparrow$0.004
	Base_{top25}	0.182	0.207	0.195
	Base_{top30}	0.192 $\uparrow$0.025	0.218 $\uparrow$0.002	0.205 $\uparrow$0.013
	$\text{DoFIT}_{\alpha=1.1}$	0.253	0.252	0.252
	$\text{DoFIT}_{\alpha=1.2}$	**0.261** $\uparrow$0.094	**0.255** $\uparrow$0.039	**0.258** $\uparrow$0.066
	$\text{DoFIT}_{\alpha=1.3}$	0.256	0.247	0.251

Experiments

Table 3: The number of parameters per round in training.

Domain	Method	Frozen	Trainable	Comm.	S-Comm.
F&G	FIT_{32qv}	6738M	4.194M	4.194M	0M
	$(\text{Base}_{top15}$ / $\text{DoFIT})_{32d}$	6738M	4.021M	4.021M	0.942M
M&G	FIT_{32qv}	6738M	4.194M	4.194M	0M
	$(\text{Base}_{top30}$ / $\text{DoFIT})_{32qv}$	6738M	4.194M	4.194M	0.983M

Table 4: Comparison with existing federated domain adaptation works.

Method	FPB		FiQA-SA		TFNS		NWGI		Avg:3		Avg:4	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
FedGP	0.837	0.829	0.760	0.806	0.789	0.786	0.625	0.626	0.795	0.807	0.753	0.762
FedGP-g	0.836	0.830	0.680	0.744	0.700	0.710	0.627	0.629	0.739	0.761	0.711	0.728
$\text{DoFIT}_{\alpha=1.0}$	0.861	0.858	0.818	0.847	0.869	0.869	0.641	0.640	0.849	0.858	0.797	0.804

Experiments

Table 5: Performance on the gradient and singular value spectrum.

Criteria	FPB		FiQA-SA		TFNS		NWGI		Avg:3		Avg:4	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DoFIT$_{\alpha=1.0}$	0.861	0.858	0.818	0.847	0.869	0.869	0.641	0.640	0.849	0.858	0.797	0.804
A-grad-top15	0.866	0.864	0.833	0.852	0.867	0.867	0.640	0.639	0.855	0.861	0.802	0.806
A-svd-top15	0.858	0.855	0.829	0.856	0.828	0.829	0.642	0.641	0.838	0.847	0.789	0.795
B-grad-top10	0.823	0.813	0.789	0.827	0.802	0.806	0.633	0.633	0.805	0.815	0.762	0.770
B-grad-top15	0.833	0.829	0.840	0.855	0.681	0.693	0.630	0.627	0.785	0.792	0.746	0.751
B-grad-top20	0.516	0.480	0.185	0.197	0.501	0.500	0.404	0.369	0.401	0.392	0.402	0.387
B-svd-top10	0.856	0.854	0.844	0.854	0.732	0.740	0.638	0.627	0.811	0.816	0.768	0.769
B-svd-top15	0.821	0.819	0.793	0.824	0.621	0.626	0.644	0.640	0.745	0.756	0.720	0.727
B-svd-top20	0.417	0.306	0.811	0.794	0.371	0.302	0.552	0.457	0.533	0.467	0.538	0.540

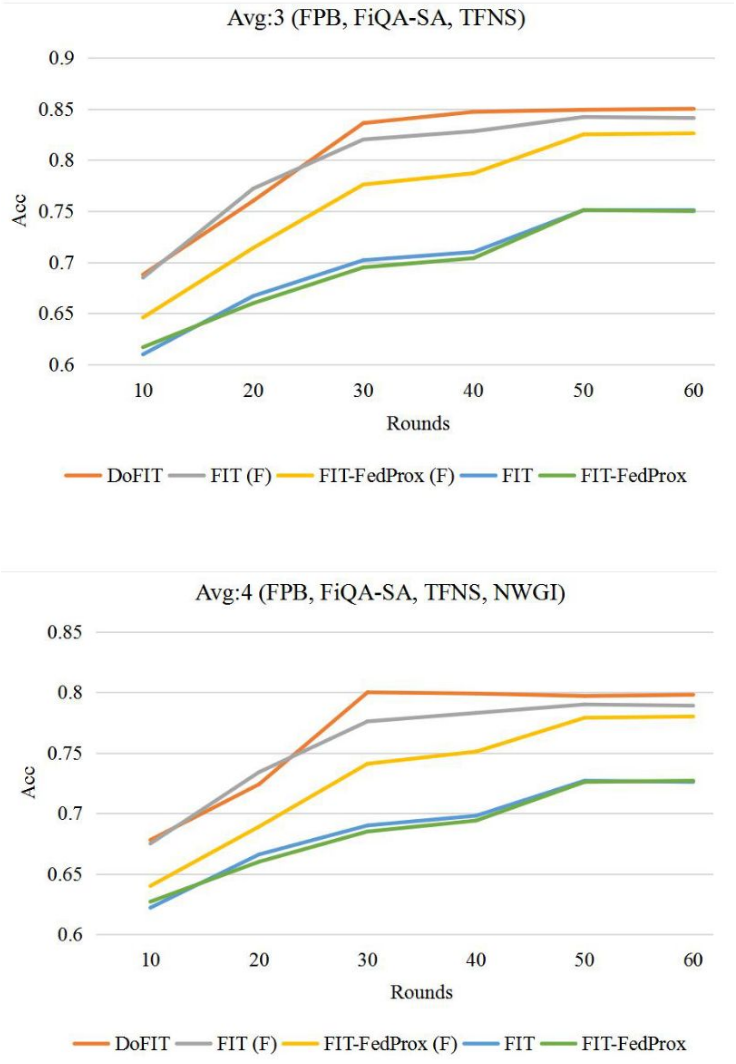


Figure 6: Comparison of average accuracy on different rounds

Experiments

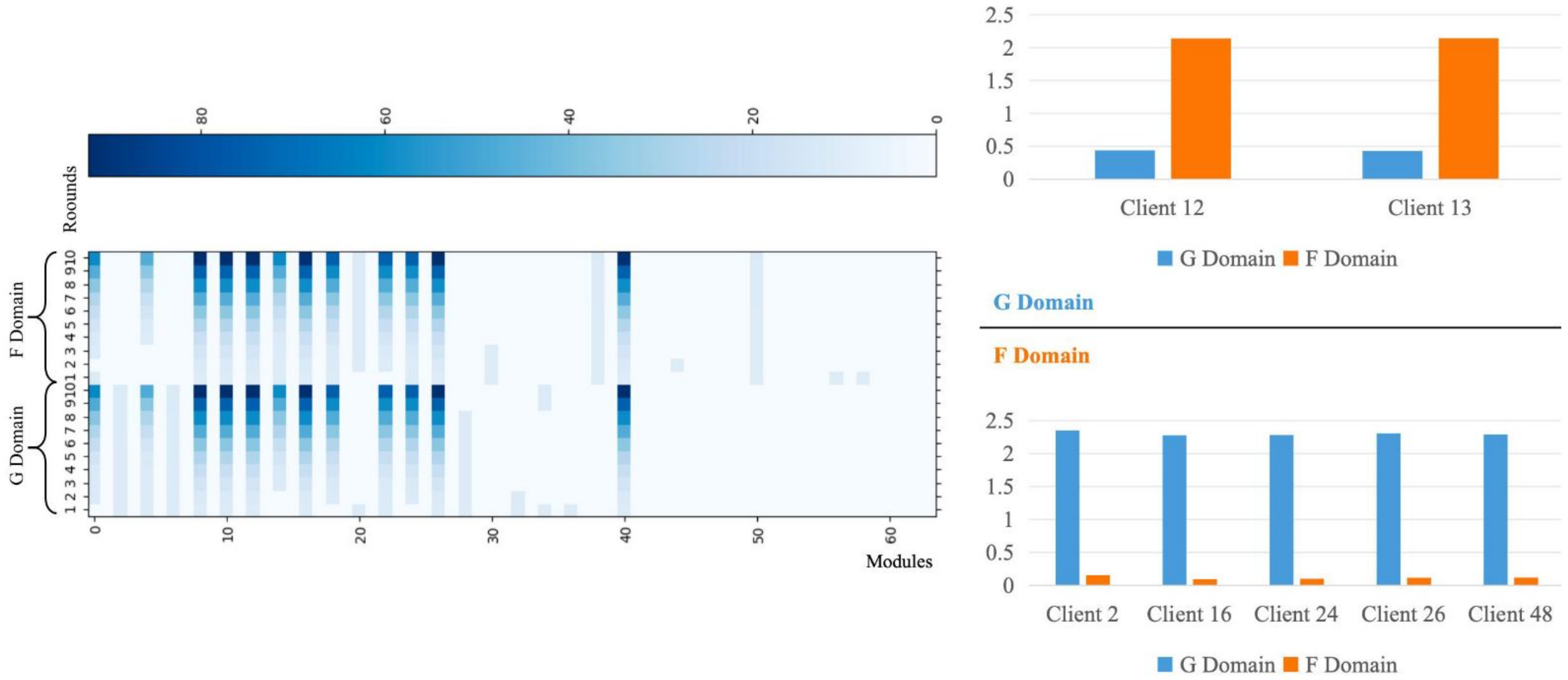


Figure 7: Modules important scores (left) and singular value spectrum (right) on F and G domains

Conclusion

- ❑ We introduced a novel Domain-aware Federated Instruction Tuning (DoFIT) framework towards collaborative training on more datasets in relevant domains for boosting the performance of individual domains.
- ❑ For aggregation, we first normally aggregate domain-specific information on the intra-domain server side, and then aggregate overlapping domain-agnostic information on the inter-domain server side, excluding the interference information.
- ❑ For initialization, we add a proximal perturbation from interdomain information to the original modules, rather than directly overwritten them.
- ❑ Comprehensive experimental results on Finance, Medical, and General domains demonstrate the effectiveness of the proposed DoFIT method, compared to conventional FIT.

Thanks!