TransFed: Epitomizing Focal Modulation in Transformer-based Federated Setup

Tajamul Ashraf[1], Fuzayil[2], Iqra Altaf Gillani[2]

[1] IIT Delhi, India [2] NIT Srinagar, India

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Transformers

Transformers utilize *self-attention* for global interactions, resilient to shifts. Self-attention mechanism is now being applied in *federated learning*, combined with the (FedAvg) algorithm for improved performance.

Focal Modulation

Given a feature map $X \in \mathbb{R}^{H\times W\times C}$, a generic encoding generates $y_i \in \mathbb{R}^C$ for each visual token x_i via interaction T with X and aggregation M over contexts. *Focal modulation [1]* refines y_i using early aggregation:

$$y_i = T_2(M(i, X), x_i). \quad (1)$$

In Equation (1), Focal Modulation is instantiated as:

$$y_i = q(x_i) \odot m(i, X), \quad (2)$$

where $q(\cdot)$ is a query projection ,$\odot$ is element-wise multiplication and $m(\cdot)$, a context aggregation.

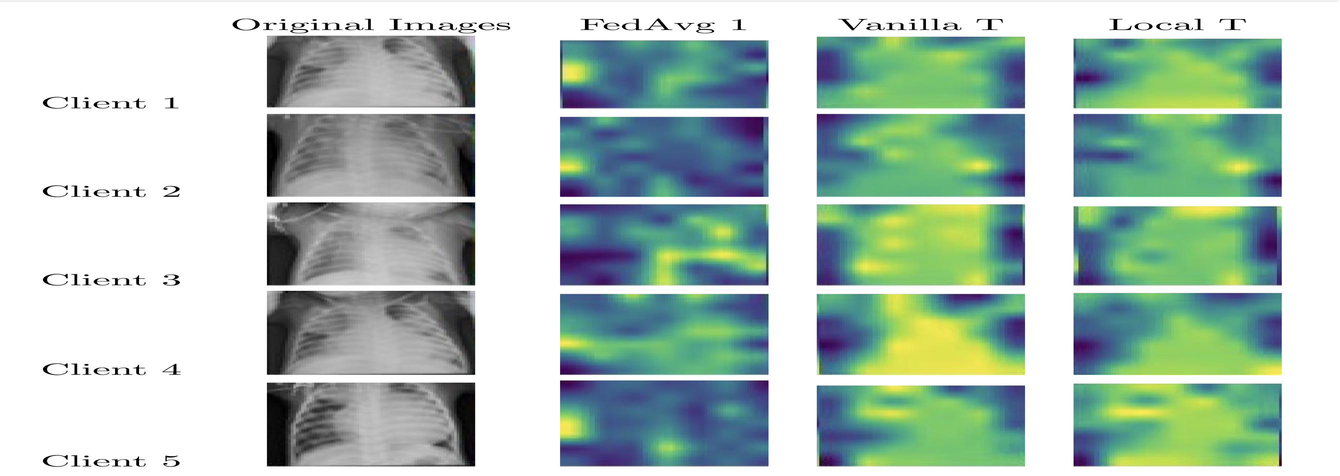

Figure 1: Comparing focal maps of Local-T, FedAvg-T, and Vanilla-T across clients, we see local training and Vanilla-T emphasizes task details, while FedAvg-T disrupts such information.

Problem statement: Mitigating data heterogeneity and building a tailored model

In a federated scenario, N clients with local datasets $D_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j=1}^{m_i}$, $1 \leq i \leq N$, contribute to a total dataset D of size $M = \sum_{i=1}^{N} m_i$. The model for client i is denoted as $f(\theta_i; \cdot)$ with parameters θ_i.

$$\arg\min_{\Theta} \sum_{i=1}^{N} \left(\frac{m_i}{S}\right) K_i(\theta_i) \quad (3)$$

Problem Characterization

TransFed uses DINO [2]. The Focal modulation mechanism operates on the queries, keys, and values, denoted as $Q = MP_Q$, $K = MP_K$, and $V = MP_V$, respectively. We concatenate these projection parameters into $P = [P_Q, P_K, P_V]$ for simplicity. By utilizing a visual feature map $X \in \mathbb{R}^{H\times P\times C}$ as the input, a standard encoding process produces a feature representation $y_i \in \mathbb{R}^C$ for each visual token (query) $Q_i \in \mathbb{R}^C$.

Proposed Solution: Custom Learning

In TransFed, a Learnable generator $h_\phi(z_i)$ at the server, parameterized by ϕ, takes a client's embedding vector $z_i \in \mathbb{R}^D$ as input. The generator produces projection parameters $P_i = h_\phi(z_i)$, decomposed into query, key, and value matrices (P_{Qi}, P_{Ki}, P_{Vi}) for focal-modulation.

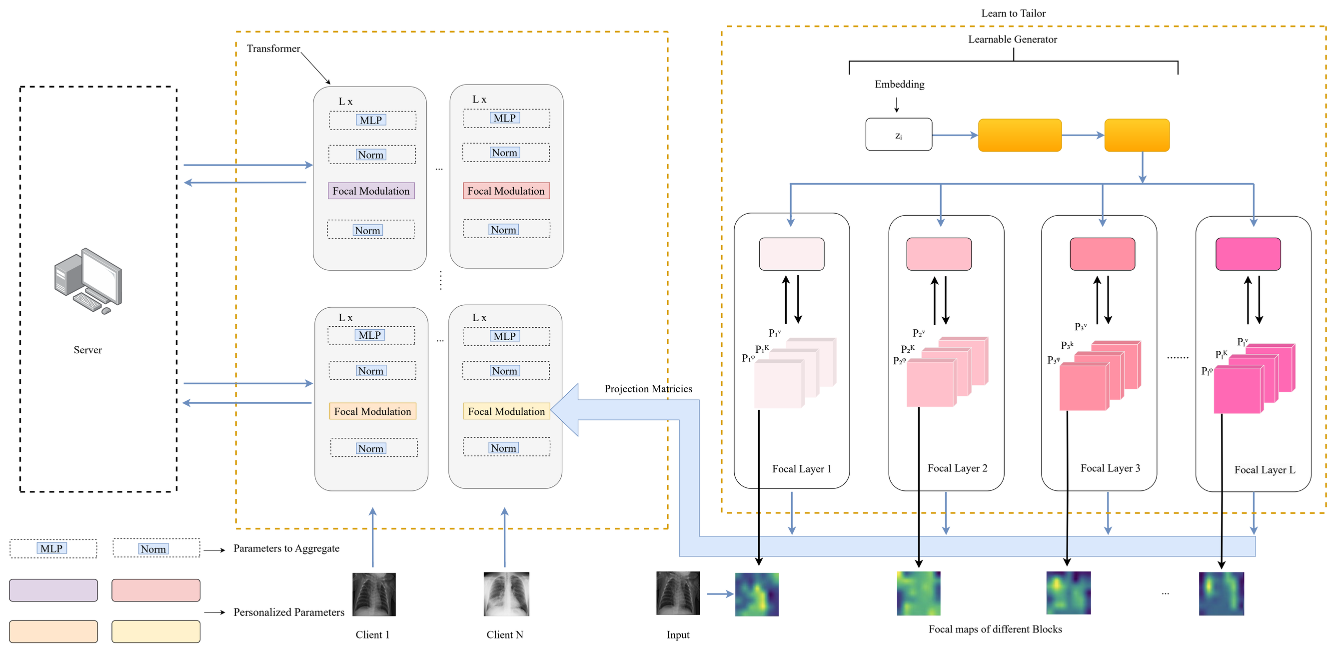

Figure 2: Combining Local Retention and Server-Based Aggregation featuring localized focal modulation layers and central parameter aggregation, fostering collaboration among clients. The 'learn-to-tailor' mechanism employs a server-based generator to create unique projection matrices in L transformer blocks, enhancing adaptability

Vanilla Tailoring

In TransFed, parameters are locally trained and aggregated on the server, akin to FedAvg. The FM layer, with parameters P_i, and other layers, with ξ, constitute the tailored model $\theta_i = (P_i, \xi)$. Local training is iterated over multiple rounds, updating the model $f(P_i^{t,k}, \bar{\xi}_i^{t,k}; \cdot)$. $P_i^{t,k}$ retains local information, and $\bar{\xi}_i^{t,k}$ aggregates across clients using

$$\bar{\xi}^t = \sum_{i=1}^{N} \left(\frac{m_i}{S}\right) \xi_i^{t,k}. \quad (4)$$

Experiments

Experiments were conducted on pneumonia benchmark datasets: Kermany [3] and RSNA [4]. Two partitioning techniques were employed to emulate non-IID scenarios.

Dataset	Task	Clients	Total Samples	Model
RSNA [4]	Image Classification	100/200	30227	FocalNet
Kermany [3]	Image Classification	100/200	5,232	FocalNet

Performance Analysis

	RSNA dataset				Kermany dataset			
# distribution	Pathological	Pathological	Beta	Beta	Pathological	Pathological	Beta	Beta
# no. of clients	100	200	100	200	100	200	100	200
Local-T	84.55±0.15	82.21±0.08	69.94±0.13	66.68±0.13	55.91±0.17	49.25±0.11	27.87±0.12	23.34±0.10
FedAvg-T	50.42±4.22	46.28±4.23	61.85±1.5	59.23±1.93	34.02±0.88	30.20±0.95	38.64±0.22	34.89±0.45
FedPer-T	89.86±0.89	89.01±0.12	79.41±0.16	77.70±0.14	67.23±0.32	61.72±0.16	37.19±0.18	29.58±0.14
pFedHN-T	82.26±0.61	77.57±0.52	71.45±0.87	68.13±0.67	53.08±0.72	39.94±0.91	33.25±0.77	29.14±0.98
Fed TP	79.75±0.22	75.46±0.11	77.25±0.69	71.13±0.84	48.61±0.45	46.05±0.47	36.63±0.98	25.13±0.35
Vanilla -T	91.83±0.27	91.28±0.12	89.23±0.78	87.77±0.37	88.67±0.54	88.23±0.11	87.74±0.12	87.26±0.85
TransFed	**92.67±0.74**	**91.34±0.86**	**88.49±0.38**	**88.16±0.33**	**89.80±0.23**	**87.73±0.74**	**87.34±0.92**	**86.98±0.64**

Table 2: TransFed's test accuracy compared with diverse transformer-based approaches in non-IID scenarios.

Analysis of Different Adapted Parts

Customized Part	RSNA		Kermany	
	Pathological	Beta	Pathological	Beta
Focal Modulation	92.67 ±0.74	88.49 ±0.38	89.80 ±0.23	87.344±0.92
MLP Layers	88.45±0.14	86.36±0.17	87.76±0.14	85.97±0.16
Normalization Layers	89.56±0.45	86.55±0.27	86.23±0.37	87.22±0.39
Encoder	82.34±0.43	83.65±0.52	83.79±0.24	83.95±0.37

Table 3: Average test accuracy of focal models with varying customized components.

Generalization to Novel Clients

Method	Personalization	Client Accuracy (%)	Convergence Time (epochs)
pFedMe	All Parameters	78.3	8
pFedHN (Embedding)	Clientwise Embedding	79.5	6
pFedHN (Hypernetwork)	Whole Hypernetwork	80.2	5
FedRod	Last Classification Layer	77.8	10
Vanilla Personalized-T	Self-Attention Projection Matrices	76.7	12
FedTP	Self Attention Layers	81.2	4
TransFed (Learnable Generator)	Focal Modulation Layers	82.6	3

Table 4: Generalization Performance Comparison on RSNA dataset.

Conclusion

Introduced TransFed, a transformer-based federated learning framework addressing FM limitations in non-IID scenarios.Enhanced FM through client tailoring via a central Learnable generator. Experimental results show TransFed outperforming with 8% and 12% increases on RSNA and Kermany, respectively, despite slower training speed.

References

[1] J. e. a. Yang, ``Focal modulation networks,'' *NuerIPS*, 2022.

[2] M. e. a. Caron, ``Emerging properties in self-supervised vision transformers,'' in *ICCV*, pp. 9650--9660, 2021.

[3] D. e. a. Kermany, ``Labeled optical coherence tomography (oct) and chest x-ray images for classification,'' *Mendeley data*, vol. 2, no. 2, p. 651, 2018.

[4] X. e. a. Wang, ``Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,'' in *CVPR*, pp. 2097--2106, 2017.