Indian Institute of Technology, New Delhi – 110016
And
National Institute of Technology, Srinagar -190006

JAN 4-8 WACV 2024
WAIKOLOA HAWAII

TransFed: A way to epitomize Focal Modulation using Transformer-based Federated Learning

Tajamul Ashraf[1] Fuzayil Bin Afzal Mir[2] Iqra Altaf Gillani[2]

Paper ID: 479

[1]Indian Institute of Technology Delhi, India
[2]National Institute of Technology Srinagar, India

Transformers in Federated Learning

- Transformers utilize **self-attention** for global interactions, resilient to shifts.

Transformers in Federated Learning

- Transformers utilize **self-attention** for global interactions, resilient to shifts.

- Transformers, with their successful self-attention mechanism, are now being applied in federated learning, combined with the Federated Averaging **(FedAvg)** algorithm for improved performance.

Transformers in Federated Learning

- Transformers utilize **self-attention** for global interactions, resilient to shifts.

- Transformers, with their successful self-attention mechanism, are now being applied in federated learning, combined with the Federated Averaging **(FedAvg)** algorithm for improved performance.

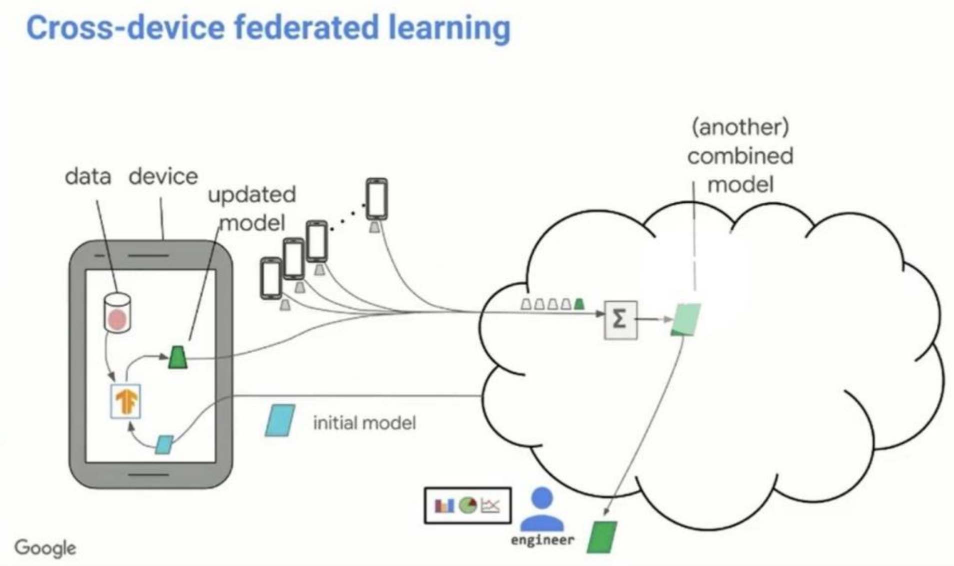

Figure 1. Illustrating Model Distribution and Combining Updates in cross-device federated learning

(Image Credits: Peter Kairouz et al.)

FocalNet Based Transformers

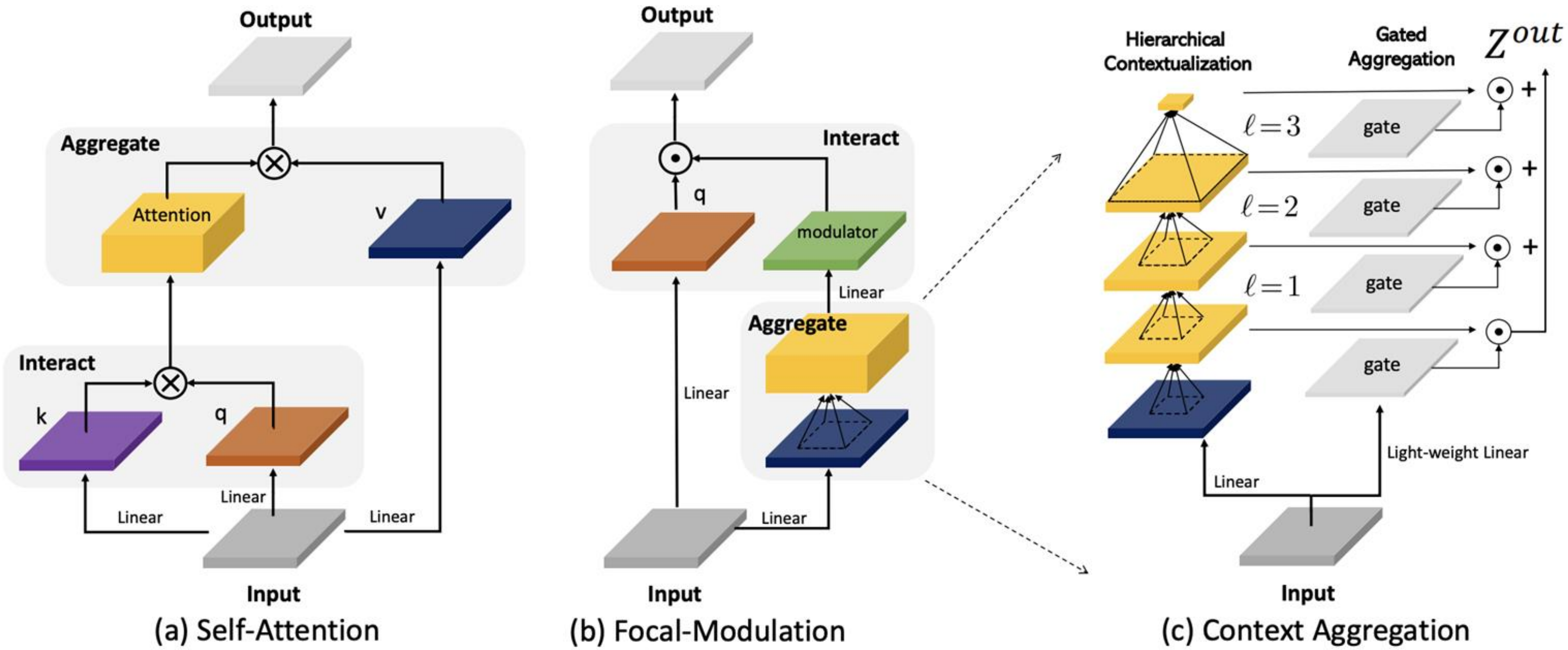

Figure 2: Left: Comparing SA (a) and Focal modulation (b) side by side. Right: Detailed illustration of context aggregation in focal modulation (c).

(Image Credits: yang et al.)

FocalNet Based Transformers

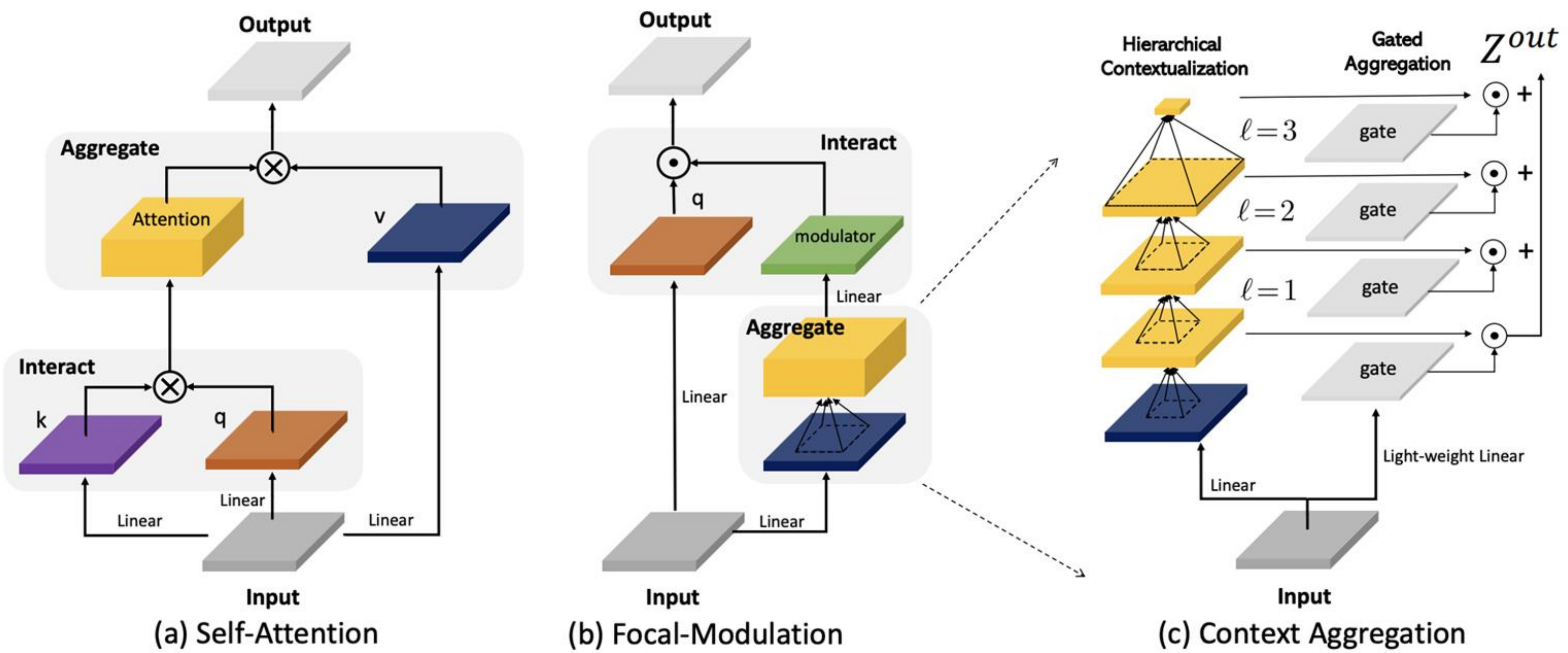

Figure 2: Left: Comparing SA (a) and focal modulation (b) side by side. Right: Detailed illustration of context aggregation in focal modulation (c).

FocalNets leverage **focal modulation** instead of self-attention, allowing for the effective modelling of interactions between tokens in visual data.

(Image Credits: yang et al.)

Comparing Focal Modulation Maps

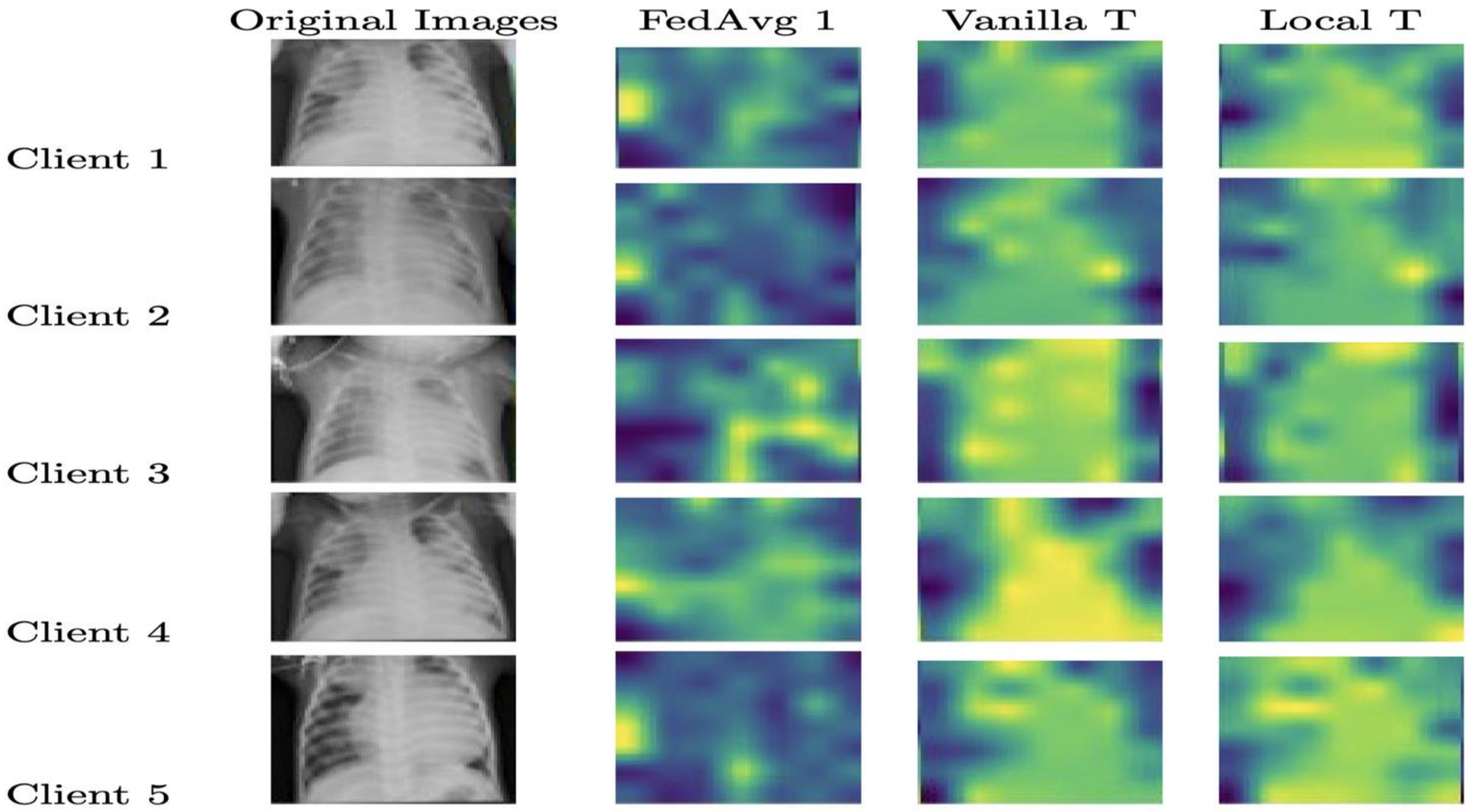

Figure 3. Comparing focal maps of Local-T, FedAvg-T, and Vanilla-T across clients, we see local training and Vanilla-T emphasize task details, while FedAvg-T disrupts such information.

Problem Statement

In a federated scenario, N clients with local datasets $Di = \{(x\ (j)\ l\ ,\ y\ (j)\ l\ )\}\ mi$ $j{=}1,\ 1 \le l \le N$, contribute to a total dataset D of size $M = \sum_{i=q}^{N} m_{1}$. The model for client l is denoted as $f(\theta_l\ ;\ \cdot)$ with parameters θ_l .

$$\arg min \sum_{l=1}^{N} \left(\frac{m_1}{S}\right) \mathrm{K}_l \theta_1$$

TransFED: Vanilla Tailoring of Focal Modulation

Our solution involves tailored focal modulation, **customizing local layers** while averaging others to preserve standard insights.

TransFED: Vanilla Tailoring of Focal Modulation

Our solution involves tailored focal modulation, **customizing local layers** while averaging others to preserve standard insights.

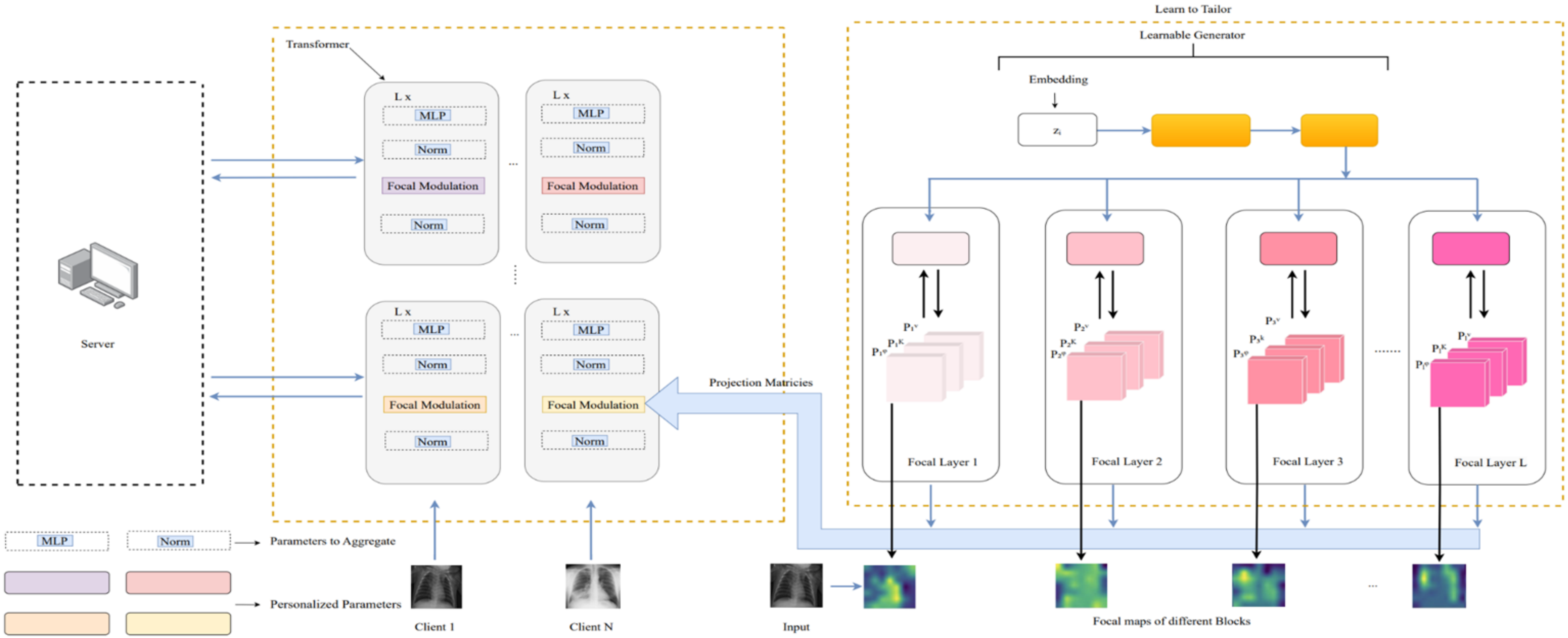

Figure 5. Comparing focal maps of Local-T, FedAvg-T, and Vanilla-T across clients, we see local training and Vanilla-T emphasize task details, while FedAvg-T disrupts such information.

Custom Learning for Focal Modulation

In TransFed, a Learnable generator $h\varphi(zi)$ at the server, parameterized by φ, takes a client's embedding vector $zi \in R\ D$ as input.

Custom Learning for Focal Modulation

In TransFed, a Learnable generator $h_\varphi(z_i)$ 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 $P_i = h_\varphi(z_i)$, decomposed into query, key, and value matrices (P^Q_i, P^K_i, P^V_i) for focal-modulation.

Custom Learning for Focal Modulation

In TransFed, a Learnable generator $h_\varphi(z_i)$ at the server, parameterized by φ, takes a client's embedding vector $z_i \in R^D$ as input.

The generator produces projection parameters $P_i = h_\varphi(z_i)$, decomposed into query, key, and value matrices (P^Q_i, P^K_i, P^V_i) for focal-modulation.

In TransFed, parameters are locally trained and aggregated on server, akin to FedAvg. The focal modulation layer, with parameters P_i, and other layers, with ξ, constitute the tailored model $\theta_i = (P_i, \xi)$.

Datasets

We conducted experiments on two widely used pneumonia benchmark datasets: **Kermany []** and **RSNA []**.

Datasets

We conducted experiments on two widely used pneumonia benchmark datasets: **Kermany []** and **RSNA []**.

We utilized two partitioning techniques to emulate *non-IID (non-identically distributed)* scenarios in our experiments.

Datasets

We conducted experiments on two widely used pneumonia benchmark datasets: **Kermany []** and **RSNA []**.

We utilized two partitioning techniques to emulate *non-IID (non-identically distributed)* scenarios in our experiments.

- **Pathological setting**
- **Symmetric Beta distribution**

Datasets

We conducted experiments on two widely used pneumonia benchmark datasets: **Kermany []** and **RSNA []**.

We utilized two partitioning techniques to emulate *non-IID (non-identically distributed)* scenarios in our experiments.

- **Pathological setting**
- **Symmetric Beta distribution**

Dataset	Task	Clients	Total Samples	Model
RSNA [31]	Image Classification	100/200	30227	FocalNet
Kermany [11]	Image Classification	100/200	5,232	FocalNet

Table 1. Datasets and Models.

Baselines

In our evaluation, we comprehensively compared **TransFed** with various federated learning algorithms.

Baselines

In our evaluation, we comprehensively compared **TransFed** with various federated learning algorithms.

- Fundamental federated algorithms: *FedAvg* and *FedProx*.

Baselines

In our evaluation, we comprehensively compared TransFed with various federated learning algorithms.

- Fundamental federated algorithms: *FedAvg* and *FedProx*.
- State-of-the-art customization algorithms: *FedPer*, *pFedMe*, *and FedTP*, as well as Vanilla-based models.

Baselines

In our evaluation, we comprehensively compared TransFed with various federated learning algorithms.

- Fundamental federated algorithms: *FedAvg* and *FedProx*.
- State-of-the-art customization algorithms: *FedPer*, *pFedMe*, *and FedTP*, as well as Vanilla-based models.

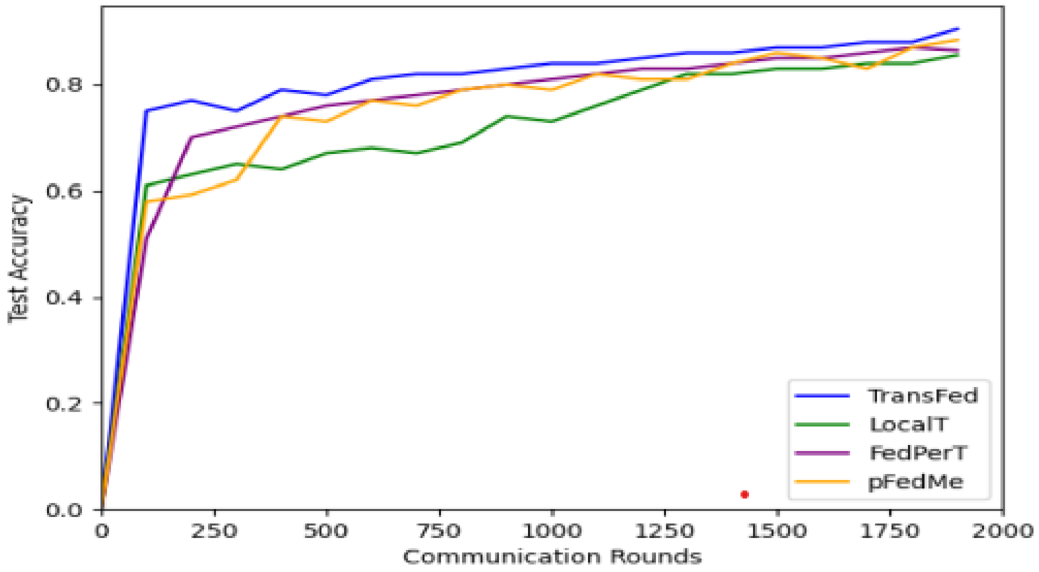

Figure 6: Test accuracy and convergence behavior of TransFed and other transformer-based methods on RSNA dataset

Performance Analysis

We conducted a comprehensive performance comparison between TransFed and several well-known federated learning methods, designed initially based on CNN backbones.

Performance Analysis

We conducted a comprehensive performance comparison between TransFed and several well-known federated learning methods, designed initially based on CNN backbones.

	RSNA dataset				Kermany dataset			
# distribution # no. of clients	Pathological 100	Pathological 200	Beta 100	Beta 200	Pathological 100	Pathological 200	Beta 100	Beta 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. The TransFed method average test accuracy is computed alongside that of multiple transformer-based approaches, encompassing different non-IID scenarios.

Analysis of Different Adapted Parts

This study examined the effects of personalizing various components of the transformer model.

Analysis of Different Adapted Parts

This study examined the effects of personalizing various components of the transformer model.

We used the same Learnable generator for all components and kept the focalnet structures consistent.

Analysis of Different Adapted Parts

This study examined the effects of personalizing various components of the transformer model.

We used the same Learnable generator for all components and kept the focalnet structures consistent.

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

We thoroughly assessed our method's capacity for generalization, comparing it with *pFedMe*, *pFedHN*, *FedRod*, and a customized-T Vanilla approach on the Kermany and RSNA datasets under the Beta configuration.

Generalization to Novel Clients

We thoroughly assessed our method's capacity for generalization, comparing it with *pFedMe*, *pFedHN*, *FedRod*, and a customized-T Vanilla approach on the Kermany and RSNA datasets under the Beta configuration.

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.

Ablation Study

We examined the impact of the number of participating clients on model performance by varying the sample rate.

Ablation Study

We examined the impact of the number of participating clients on model performance by varying the sample rate.

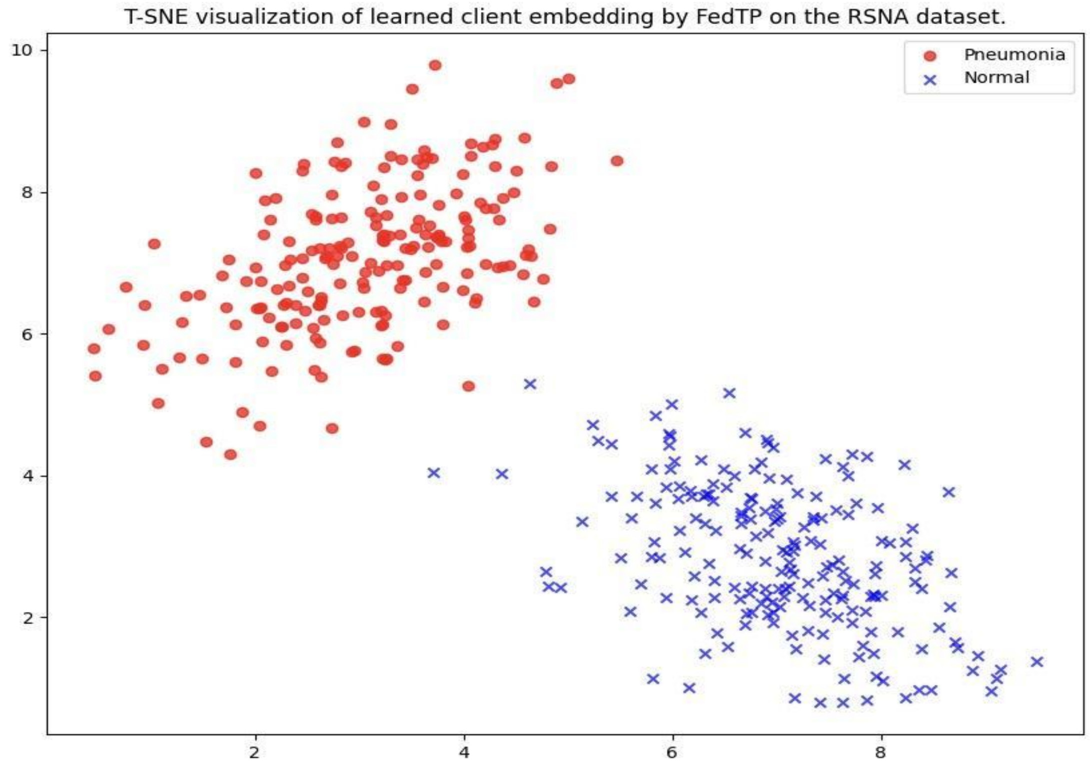

Figure 6. Visualization of Client Embeddings Learned by TransFed using **t-SNE** on the RSNA Dataset.

TransFed on CIFAR 10 and CIFAR 100

In our comprehensive comparative analysis of the experimental outcomes, our novel TransFed model emerges as a standout performer in direct contrast to the state-of-the-art benchmark methods

TransFed on CIFAR 10 and CIFAR 100

In our comprehensive comparative analysis of the experimental outcomes, our novel TransFed model emerges as a standout performer in direct contrast to the state-of-the-art benchmark methods

	Cifar 10				Cifar 100			
settings	Pathological		Dirichlet		Pathological		Dirichlet	
Client	50	100	50	100	50	100	50	100
FedAvg [8]	47.79±4.48	44.12±3.10	56.59±0.91	57.52±1.01	15.71±0.35	14.59±0.40	18.16±0.58	20.34±1.34
FedProx [6]*	50.81±2.94	57.38±1.08	58.51±0.65	56.46±0.66	19.39±0.63	21.32±0.71	19.18±0.30	19.40±1.76
FedPer [2]*	83.39±0.47	80.99±0.71	77.99±0.02	74.21±0.07	48.32±1.46	42.08±0.18	22.60±0.59	20.06±0.26
pFedMe [9] *	86.09±0.32	85.23±0.58	76.29±0.44	74.83±0.28	49.09±1.10	45.57±1.02	31.60±0.46	25.43±0.52
FedBN [7]*	87.45±0.95	86.71±0.56	74.63±0.60	75.41±0.37	50.01±0.59	48.37±0.56	28.81±0.50	28.70±0.46
pFedHN [4]*	88.38±0.29	87.97±0.70	71.79±0.57	68.36±0.86	59.48±0.67	53.24±0.31	34.05±0.41	29.87±0.69
pFedGP [1]*	89.20±0.30	88.80±0.20	---	---	63.30±0.10	61.30±0.20	---	---
FedRoD [3]*	89.87±0.03	89.05±0.04	75.01±0.09	73.99±0.09	63.30±0.10	61.30±0.20	---	---
FedTP [5]	90.31±0.26	88.39±0.14	81.24±2.17	80.27±0.28	68.05±0.24	63.76±0.39	46.35±0.29	43.74±0.39
TransFed (Ours)	**93.47±0.75**	**91.85±0.39**	**82.89±0.75**	**79.75±0.15**	**71.96±0.54**	**68.11±0.39**	**51.75±0.12**	**44.33±0.74**

Table 5. Results OF FedTP and other Benchmark methods on Image datasets with different Non-IID settings.

Conclusion

- We introduced **TransFed,** a transformer-based federated learning framework that addresses the limitations of Focal Modulation in non-IID scenarios.

Conclusion

- We introduced **TransFed**, a transformer-based federated learning framework that addresses the limitations of Focal Modulation in non-IID scenarios.

- TransFed enhances the performance of Focal Modulation by tailoring it to each client through the use of a **central Learnable generator**.

Conclusion

- We introduced TransFed, a transformer-based federated learning framework that addresses the limitations of Focal Modulation in non-IID scenarios.

- TransFed enhances the performance of Focal Modulation by tailoring it to each client through the use of a central Learnable generator.

- Experimental results demonstrate TransFed's superiority in non-IID contexts, with an increase in **8%** and **12%** on RSNA and Kermany respectively.

References

1. Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. Advances in Neural Information Processing Systems, 35:4203–4217, 2022.

1. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Y Arcas. Communication efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017. 1, 6

1. Hongxia Li, Zhongyi Cai, Jingya Wang, Jiangnan Tang, Weiping Ding, Chin-Teng Lin, and Ye Shi. FedTP: Federated learning by transformer personalization. IEEE Transactions on Neural Networks and Learning Systems, 2023.

1. Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. arXiv preprint arXiv:1912.00818, 2019.

1. David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. arXiv preprint arXiv:1609.09106, 2016.

Indian Institute of Technology, New Delhi – 110016
And
National Institute of Technology Srinagar

JAN 4-8 WACV 2024
WAIKOLOA HAWAII

Thank You

tajamul@sit.iitd.ac.in
mfuzayil@gmail.com
iqraaltaf@nitsri.ac.in