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

Tajamul Ashraf
IIT Delhi
New Delhi, India
www.tajamulashraf.com

Fuzayil Bin Afzal Mir
NIT Srinagar
Srinagar, India
mfuzayil@gmail.com

Iqra Altaf Gillani
NIT Srinagar
Srinagar, India
iqraaltaf@nitsri.ac.in

Abstract

Federated learning has emerged as a promising paradigm for collaborative machine learning, enabling multiple clients to train a model while preserving data privacy jointly. Tailored federated learning takes this concept further by accommodating client heterogeneity and facilitating the learning of personalized models. While the utilization of transformers within federated learning has attracted significant interest, there remains a need to investigate the effects of federated learning algorithms on the latest focal modulation-based transformers. In this paper, we investigate this relationship and uncover the detrimental effects of federated averaging (FedAvg) algorithms on Focal Modulation, particularly in scenarios with heterogeneous data. To address this challenge, we propose TransFed, a novel transformer-based federated learning framework that not only aggregates model parameters but also learns tailored Focal Modulation for each client. Instead of employing a conventional customization mechanism that maintains client-specific focal modulation layers locally, we introduce a learn-to-tailor approach that fosters client collaboration, enhancing scalability and adaptation in TransFed. Our method incorporates a hyper network on the server, responsible for learning personalized projection matrices for the focal modulation layers. This enables the generation of client-specific keys, values, and queries.Furthermore, we provide an analysis of adaptation bounds for TransFed using the learn-to-customize mechanism. Through intensive experiments on datasets related to pneumonia classification, we demonstrate that TransFed, in combination with the learn-to-tailor approach, achieves superior performance in scenarios with non-IID data distributions, surpassing existing methods. Overall, TransFed paves the way for leveraging focal Modulation in federated learning, advancing the capabilities of focal modulated transformer models in decentralized environments.

1. Introduction

Federated learning is a concept aimed at training a collective global model by employing data from numerous clients, all while upholding the confidentiality of the data [21]. To address concerns regarding data privacy and communication overhead, each client trains its local model and shares only the model updates usually weights and parameters with the server. However, learning a single global model may be ineffective when dealing with heterogeneous data and system variations across different clients. Tailored federated learning has emerged as an extension of the federated learning approach to address this challenge. This paradigm centers around the learning of tailored models, diverging from using a single global model while simultaneously harnessing the advantages of collaborative training [26], [25]. Numerous methods have emerged to tackle the issue presented by non-IID (non-independent and identically distributed) data distribution among clients. The majority of current federated learning frameworks predominantly utilize CNNs, which demonstrate susceptibility to variations in diverse data [8]. Transformers [28], on the other hand, utilize a self-attention mechanism to capture global interactions among inputs, rendering them more resilient to distribution shifts and data heterogeneity [24].

Inspired by the success of self-attention, recent research has explored the use of transformers as the underlying network architecture for federated learning, in combination with the fundamental federated averaging (FedAvg) algorithm [21]. Furthermore, a novel architecture based on focal modulation has been proposed to explore alternative approaches for modeling input-dependent long-range interactions [32]. Focal modulation networks leverage a focal modulation module instead of self-attention, allowing for the effective modeling of interactions between tokens in visual data. While early experiments have demonstrated encouraging outcomes in transformers based federated learning, a comprehensive investigation into the influence of federated learning algorithms on focal modulation remains

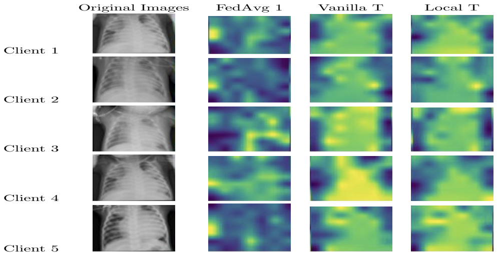

Figure 1. Comparing focal maps of Local-T, FedAvg-T, and Vanilla-T across clients, we see local training and Vanilla-T emphasize task details, while FedAvg-T disrupts such information.

pending. Such algorithms might restrict the capabilities of focalnet based transformers [32] in the context of federated learning. Considering the potential of focal-modulated based federated learning, it is imperative to conduct further research to explore this topic comprehensively. Recent studies have highlighted the crucial role of focal modulated layers in transformers [32], [30], [37], emphasizing their significance compared to other architectures. Expanding on this observation, we conducted an inquiry into the impact of focal modulation on federated learning. The experimentation involved several model variations: 1) Local-T, entailing the training of individual ViT models on each client; 2) FedAvg-T, implementing the FedAvg algorithm for the training of a global focalnet model; and 3) Vanilla Tailored-T, which retained local focal modulation while utilizing FedAvg for server-side aggregation of other parameters. To experiment, we sampled a set of images from the RSNA pneumonia dataset [31]. The public dataset is a well-known and widely used in medical imaging. It contains an extensive collection of chest X-ray images with labels indicating the presence or absence of pneumonia [31], consisting of two classes (normal and pneumonia), across five clients, as incorporating data from five clients helped capture a broader range of variations and patterns in the dataset. As we increase the number of clients, the communication required between clients and the central server also escalates. The generation of attention maps for analysis was carried out using the Attention Rollout technique, as outlined in prior work [1]. The visual representation in Figure 1 showcases the attention maps produced through the mentioned approaches, accompanying the original images. Our assessment indicates that both the Local-T and Vanilla Tailored-T models adeptly identified crucial information in the images, indicated by the highlighted regions in yellow. At the same time, FedAvg-T was unsuccessful in generating meaningful focal modulation maps [32]. Additionally, the scalability of focal modulated layers is limited as their number increases linearly with the number of clients. There are limitations in adapting tailored focal Modulation to new clients, as it necessitates retraining all the focal modulation layers. In light of these constraints, we have introduced a novel federated learning framework called TransFed, which employs a learn-to-tailor concept replacing the conventional approach of customization based on the focalnet model.

In the TransFed framework, a Learnable generator is trained on the server to generate matrix projections within the focal modulation layers. These matrices enable the creation of client-specific queries, keys, and values while aggregating and sharing other model parameters. Utilizing the learn-to-tailor mechanism for focal modulation layers via the Learnable generator, we can efficiently distribute parameters across clients and generate customized focal modulation layers. TransFed not only achieves exceptional accuracy but also demonstrates scalability as the number of clients increases and showcases strong adaptability to new clients.

The primary contributions are summarized as follows:

- We introduce TransFed, an innovative federated learning framework built upon Focal modulation architecture. TransFed directly addresses the limitations of FedAvg when applied to focal modulation in heterogeneous data scenarios. By facilitating the customization of focal modulation for individual clients, TransFed significantly improves performance within the context of tailored federated learning.

- Our proposal introduces a learn-to-tailor concept to enhance the utilization of client cooperation in the tailored layers. This mechanism aims to improve the scalability and adaptation capabilities of TransFed.

- Comprehensive experiments were carried out on benchmark datasets for pneumonia, encompassing various scenarios of non-IID data distribution. The results of these experiments unequivocally demonstrate that TransFed outperforms a wide range of benchmark methods in tailored federated learning, specifically in image-based tasks, establishing TransFed as the state-of-the-art solution in this domain.

The rest of the article is organized as follows: Section 2 provides a brief overview of prior work on tailored federated learning, transformers, and Learnable generators. Section 3 outlines the tailored federated learning formulation with transformers, including two customization methods: vanilla and learn-to-tailor. Section 4 presents experimental results and discussions, thoroughly analyzing the findings. In Section 5, TransFed's limitations are discussed, concluding with insights into future directions.

2. Related Work

Various approaches have been proposed to address the heterogeneity among clients in tailored federated learning

[35], [34], [33]. One approach involves fine-tuning the global model using client's local datasets to obtain tailored parameters [29], [19], [7]. Another strategy is to incorporate proximal regularization terms to handle client drift issues resulting from statistical heterogeneity, as demonstrated by FedProx [15], pFedMe [27], and Ditto [14]. Conversely, FedAlign [22] addresses the data heterogeneity challenge in federated learning from the perspective of local learning generality rather than proximal restriction.

Knn-Per [20] proposes a hybrid model that combines two existing models. The local model in this approach utilizes a k-nearest neighbors method, which requires storing all the features of the samples. Some methods, like FedMD [12] and FedGen [36], employ knowledge distillation from a global teacher model to improve the flexibility of tailored model architectures on client devices. These methods enable clients to obtain more robust tailored models. To address scenarios where data distribution varies among clients or inherent partitions exist, clustered federated learning (CFL) [25], privacy-preserving federated adaptation (PFA) [17], and FedAMP [10] utilize clustering approaches to train federated learning models for homogenous client groups. This ensures that the training process is better suited for such cases. FedTP [13] delves deeper, particularly into the ViT [6]; they only personalized the attention maps of the transformer model on each client, while the remaining parts remain as shared components. It adopts a parameter decoupling approach for cross-attention to accommodate data disparities. However, this strategy may introduce optimization misalignment between attention and the original head upon re-coupling, particularly in scenarios involving heterogeneous data distributions. Other methods, such as FedPer [2], focus on learning tailored classifier heads locally while sharing base layers.

In contrast to these existing works, our proposed TransFed framework utilizes a learn-to-customize mechanism to train tailored focal-modulation layers within a transformer. This mechanism effectively addresses client data heterogeneity, offering a novel approach to tailored federated learning. The transformer model [28] was initially developed to enhance the efficiency of machine translation tasks. Researchers have explored the applicability of transformer models to vision tasks, resulting in developments like the Vision Transformer (ViT) [6], DeTR [3], and DINO-based models [4]; DINO is particularly noteworthy. As an early exploration of using transformers in federated learning, Qu et al. [23] conducted extensive empirical experiments and demonstrated that transformers are more suitable than convolutional neural networks (CNNs) in federated learning scenarios with heterogeneous data distributions. Our TransFed framework aims to harness the full potential of the transformer architecture by training a tailored transformer for each client. The mechanism of focal modulation captures the diversity of data using a trainable generator located on the server. This generator produces projection parameters within the focal modulation layers.

Recently, Jianwei et al. [32] introduced focal modulation networks that incorporate a focal modulation mechanism to model token interactions in vision tasks, entirely replacing traditional focal modulation techniques. Crucial to the setup are learnable generators [9], which entail neural networks with the ability to produce weights for a larger target network. This is achieved using a trainable embedding vector as an input. In the context of tailored federated learning, pFedHN was the first method to utilize a Learnable generator [9], where the server's Learnable generator generates tailored weights for local convolutional neural networks (CNNs) on each client. In a similar vein, pFedLA [18] utilizes a trainable generator situated on the server to generate combined weights pertaining to individual layers of the local model across various clients. In contrast, FedRoD [5] employs a local Learnable generator that generates customized client predictors, considering additional inputs such as the client's class distributions. It is worth noting that all these Learnable generator-based methods are specifically designed for CNN architectures. However, TransFed distinguishes itself by utilizing a Learnable generator that generates projection matrices within the focal-modulation layers of a transformer. This unique approach enables the generation of client-specific queries, keys, and values.

3. Federated Learning by Tailored Focal Modulation

This section introduces our TransFed framework, which is specifically designed to address heterogeneity and produce personalized, high-quality models for individual clients. The key focus of TransFed is to learn tailored focal modulation layers for each client, enabling effective customization within the framework.

3.1. Problem Statement

In the context of visual tasks, the TransFed framework incorporates the utilization of a traditional DINO model [4]. The initial step in processing the input sequence S, which has a fixed length l, involves partitioning the images into a sequential format during the image preprocessing phase of the focalnet. Subsequently, this sequential representation is converted into an embedding matrix M, with dimensions $\mathbb{R}^{n \times m}$. 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) $Q_i \in \mathbb{R}^C$.

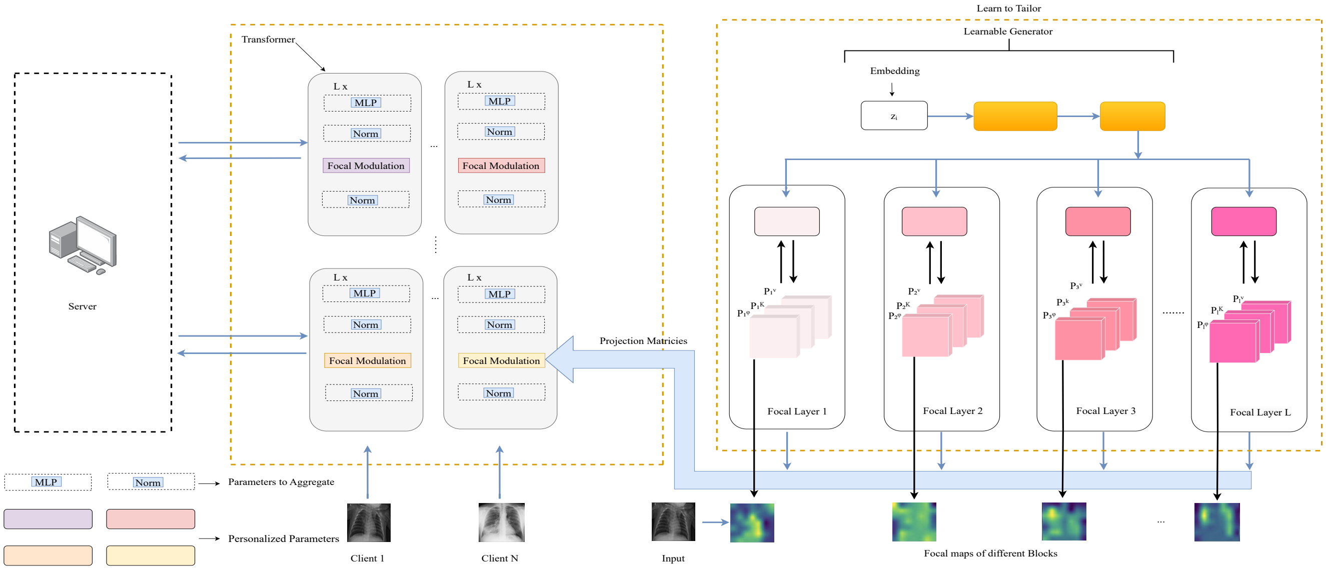

Figure 2. TransFed Overview: Combining Local Retention and Server-Based Aggregation. The architecture features localized focal modulation layers and central parameter aggregation, fostering collaboration among clients. Additionally, the 'learn-to-tailor' mechanism employs a server-based generator to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

This generation is accomplished through the token's interaction with its surroundings, including neighboring tokens, and the aggregation of information across contexts. The process involves the interaction function τ and the aggregation function A. Consequently, the refined representation y_i is obtained by combining the aggregated context features, obtained through the function A at each location i, with the query Q_i through the interaction function τ.

Focal Modulation generates refined representation y_i using an early aggregation procedure formulated as:

$$y_i = \tau(A(i, X), Q_i) \tag{1}$$

where the context features are first aggregated using A at each location i, then the query interacts with the aggregated feature based on τ to form y_i.

For the purpose of emulating a federated scenario, we examine a collection of N clients designated by $[N]$, with each specific client i holding its own local dataset$D_i = (x_i^{(j)}, y_i^{(j)})_{j=1}^{m_i}$ $(1 \leqslant i \leqslant N)$, consisting of m_i samples drawn from a distinct data distribution P_i. The total dataset is denoted as $D = \bigcup_{i \in [N]} D_i$, with a total size of $S = \sum_{i=1}^{N} m_i$.

The customized model associated with client i, defined by the parameters θ_i, is denoted as $f(\theta_i; \cdot)$. The optimization objective is defined as follows:

$$\arg\min_{\Theta} \sum_{i=1}^{N} \left(\frac{m_i}{S}\right) K_i(\theta_i) \tag{2}$$

where $K_i(\theta_i) = \mathbb{E}_{(x(j)i, y_{(j)}i) \in D_i} [l(f(\theta_i; (x_i^{(j)}), (y_i^{(j)})].$

Here,$\Theta = \theta_{i}{}_{i=1}^{N}$ represents the set of tailored parameters for each client, and $l(\cdot, \cdot)$ denotes the per-sample loss function that is common to all clients. The selection of the loss function for a specific task, whether it is a mean square error or cross-entropy loss, is contingent upon the nature of the task.

3.2. Vanilla Tailoring of Focal Modulation

Federated learning's popularity stems from global insights via focal modulation layers. But using TransFed on client layers can harm performance with diverse data. To address this challenge, our solution involves tailored focal modulation. This approach entails customizing certain local layers while averaging other layers to maintain standard insights. The diagram on the left side of the Figure 2 illustrates the fundamental configuration of the vanilla focal modulation customization.

In TransFed, parameters are locally trained and aggregated on the server, similar to FedAvg. The focal-modulation layer has projection parameters P_i, while other layers have parameters ξ. The tailored model, denoted as $\theta_i = (P_i, \xi)$, undergoes local training. This process is repeated for multiple communication rounds. Resulting in the updated model $f(P_i^{t,k}, \bar{\xi}_i^{t,k}; \cdot)$, where $P_i^{t,k}$ is retained locally to store the tailored information of client i, and $\bar{\xi}_i^{t,k}$ is aggregated across the clients using Equation (2):

$$\bar{\xi}^t = \sum_{i=1}^{N} \left(\frac{m_i}{S}\right) \xi_i^{t,k} \tag{3}$$

Consequently, the objective function of TransFed, as de-

Algorithm 1 TransFed: Tailored Focal Modulation

```
Require: $C$ - number of communication rounds, $L$ - number of local epochs, $\alpha$ - learning rate of local update, $\beta$ - learning rate for global update.
  Initialize parameters $\xi^0$, $z^{0\,i}$, and $\phi^0$;
  for $c = 1$ to $C$ do
     Sample the set of clients $C_c \subset \{1, \dots, N\}$;
     for $i$ in $C_c$ do
        $\xi^{c,0} = \bar{\xi}^{c-1}$;
        $P_i^{c,0} = h(\phi_{c-1}, z_i^{c-1})$;
        $\theta_i^{c,0} = \{P_i^{c,0}, \xi^{c,0}\}$;
        for $L = 1$ to $L$ do
           Sample mini-batch $B_i \in D_i$;
           $\theta_i^{c,k+1} \leftarrow \theta_i^{c,k} - \alpha \nabla_{\theta_i} L_i(\theta^i_{c,k}; B_i)$;
        end for
        $\Delta P_i = P^i_{c,L} - P^i_{c,0}$;
     end for
     $\bar{\xi}_t = \sum_{i \in C^c} \frac{m_i}{S} \xi^{c,L}$;
     $\phi_c = \phi_{c-1} - \beta \sum_{i \in C_c} \frac{m_i}{M} \nabla_\phi P_i^C \Delta P_i$;
     $z_c^i = z_{c-1}^i - \beta \sum_{i \in C_C} \frac{m_i}{M} \nabla_{z_i} P_i^C \Delta P_i$;
  end for
  return $\bar{\xi}^t$, $\phi^t$, and $z^t$
```

rived from Equation (1), is to minimize the following loss:

$$\arg\min_\theta \sum_{i=1}^{N} \left(\frac{m_i}{S}\right) K_i(P_i, \xi) \tag{4}$$

where

$$K_i(P_i, \xi) = \sum_{i=1}^{N} (\frac{m_i}{M} \mathbb{E}_{(x_i^j, y_i^j) \in D_i} l(f(P_i, \xi; x_i^{(j)}), y_i^{(j)}) \tag{5}$$

While the vanilla customization procedure generates tailored focal-modulation layers through local training, it overlooks the potential inherent client relationships, leading to suboptimal tailored models. Moreover, the scalability of tailored focal-modulation layers becomes an issue as the number of clients grows linearly. Additionally, the adaptation capability of tailored focal Modulation is limited, requiring retraining when novel clients are introduced to obtain specific focal-modulation layers for them.

3.3. Custom Learning for Focal Modulation

This section introduces TransFed, a framework that incorporates a learn-to-tailor approach to augment the existing vanilla customization mechanism for focal Modulation.

Algorithm 1 presents the TransFed process for parameter updates in a federated learning scenario. It spans communication rounds (C) and local epochs (L), iterating through clients to locally update model parameters (θ) using mini-batches. Global parameters ϕ and z_i are also updated collectively, yielding refined global parameters $\bar{\xi}^t$, ϕ^t, and z^t. These enhancements encourage effective collaboration among clients while retaining individual data characteristics. In the TransFed approach, a Learnable generator [9] is integrated into the server's functionality, generating projection matrices intended for the focal-modulation layers of individual clients (as illustrated on the right side of Figure 2). This design facilitates effective sharing of parameters among the clients.

The Learnable generator at the server, denoted as $h(\phi; z_i)$ and parameterized by ϕ, takes as input a learnable embedding vector $z_i \in \mathbb{R}^D$ associated with client i, which can either be a client-specific embedding or a fixed vector. We implement the Learnable generator using simple fully connected layers, where each transformer block's last layer is unique. The z_i vector, the Learnable generator produces the projection parameters $P_i = h(\phi; z_i)$ for client i, which are decomposed into the query, key, and value projection matrices for the focal-modulation mechanism, denoted as $P_i = [P_{Qi}, P_{Ki}, P_{Vi}]$.

This approach enables the Learnable generator to learn a set of projection parameters $P_i = h(\phi; z_i) | 1 \leqslant i \leqslant N$ for tailored focal Modulation. Consequently, the tailored model is represented as $f(P_i, \xi; \cdot) = f(h(z_i; \phi), \xi; \cdot)$, and loss function is updated as follows:

$$= \sum_{i=1}^{N} \left(\tfrac{m_i}{M}\right) \mathbb{E}_{(x_i^j, y_i^j) \in D_i} l(f(h(\phi; z_i), \xi; x_i^{(j)}), y_i^{(j)}) \tag{6}$$

This updated loss function computes the training loss for client i by applying the Learnable generator-generated projection parameters $P_i = h(\phi; z_i)$ alongside the standard parameters ξ to the tailored model. The update mechanism within each epoch is represented by the variable k, and the local model parameter θ_i is subjected to updates through stochastic gradient descent (SGD), as defined by the following equation:

$$\theta_i^k \leftarrow \theta_i^{k-1} - \alpha \nabla_{\theta_i} K_i(\theta_i^{k-1}; B_i) \tag{7}$$

where B_i represents a mini-batch extracted from D_i.

Representing the collection of selected clients in each round t as C_t. The gradients of ϕ and z_i can be obtained from Equation (6) using the chain rule:

$$\nabla_\phi K_i = \sum_{i \in C^t} \left(\frac{m_i}{M}\right) \nabla_\phi P_i^T \triangle_{p_i} \tag{8}$$

$$\nabla_{z_i} K_i = \sum_{i \in C^t} \left(\frac{m_i}{M}\right) \nabla_{z_i} p_i^T \triangle_{P_i} \tag{9}$$

where $\triangle P_i = P_{Ki} - P_{Qi}$ represents the change in projection parameters after K epochs of local updates.

During communication round t, updates are applied to the Learnable generator parameter ϕ and the client embedding z_i through the utilization of computed gradients:

$$\phi^t = \phi^{t-1} - \beta \nabla_\phi K_i^{(t-1)} \tag{10}$$

$$z_i^t = z_i^{t-1} - \beta \nabla_{z_i} K_i^{(t-1)} \tag{11}$$

Contrasted with the standard customization method, the learn-to-tailor approach within TransFed brings forth a range of benefits. Firstly, it achieves effective parameter sharing across clients while harnessing the focal modulation mechanism's potency in federated learning. Secondly, its scalability accompanies an expanding client base, owing to the shared Learnable generator with personalized embedding vectors driving the focal modulation layer's generation. Lastly, its adaptability extends to new clients whose data remains unseen during training. The initial and final aspects will undergo validation in Section IV. The middle aspect is substantiated by a comparison of parameter counts between learn-to-customize and standard customization. The Learnable generator, adopting an MLP architecture, comprises parameters roughly equivalent to $D_h \times D_s$, where D_h and D_s denote the dimensions of the hidden layers within the Learnable generator and the self-attention projection parameters, respectively. In contrast, the cumulative self-attention projection parameters in the standard customization rise linearly with the client count, i.e., $N \times D_s$. With a substantial client count ($N > D_h$), learn-to-customize for focal modulation consumes fewer resources.

4. Experiments

This section introduces the experimental configuration, assesses the performance of our proposed model, and conducts comparisons with several baseline methods across diverse learning scenarios. We introduce the benchmarks, non-IID settings, model architectures used in our experiments, and relevant implementation details.

4.1. Experimental Setup

4.1.1 Baselines

In our evaluation, we comprehensively compared TransFed with various federated learning algorithms. We compared TransFed against fundamental federated algorithms, including FedAvg [21] and FedProx [15]. Additionally, we evaluated its performance against state-of-the-art customization algorithms, including FedPer [2], pFedMe [27], and FedTP [13], as well as Vanilla-based models. By including these various algorithms in our comparison, we aimed to assess the effectiveness and superiority of TransFed in achieving customized and efficient federated learning. The selected algorithms represent a range of approaches that tackle different aspects of customization in federated learning, allowing us to evaluate TransFed's performance with basic and advanced techniques. This comprehensive evaluation provides valuable insights into the relative strengths and weaknesses of TransFed compared to existing state-of-the-art algorithms, further highlighting its potential as an advanced customization approach in the domain of federated learning.

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.

4.1.2 Non-IID Settings of Pnemunia Datasets

We conducted experiments on two widely used pneumonia benchmark datasets: Kermany [11] and RSNA [31].We utilized two partitioning techniques to emulate non-IID (non-identically distributed) scenarios in our experiments.

The first strategy, the *Pathological* setting, involved randomly assigning classes to each client in both the Kermany and RSNA datasets. In this setting, the sample rate for class c on client i was determined by $a_{i,c}/\sum_j a_{j,c}$, where $a_{i,c}$ was randomly generated from a uniform distribution $U(0.4, 0.6)$. The second approach consisted of generating a federated version of the datasets by randomly partitioning samples with identical labels across clients. This partitioning was done using a symmetric *Beta distribution* with a parameter of $\alpha = 0.3$. By applying the Beta distribution, we divided the samples among the clients in a federated manner, ensuring a diverse distribution of samples for training.

To enhance the realism of the local datasets within the Kermany dataset, we employed a two-stage Pachinko allocation method. This method first generated a Beta distribution with a parameter of $\alpha = 0.4$ over coarse labels for each client. Subsequently, a Beta distribution with a parameter of $\beta = 10$ was generated over the acceptable labels corresponding to the coarse labels. The class distribution and the allocation of classes in the training and test sets were kept consistent for both the coarse and fine label partitions among clients. Table 1 provides a summary of the datasets, their associated tasks, as well as the counts of clients and models involved.

4.1.3 TransFed Setup

Following the experimental setup described in pFedHN, we performed experiments using TransFed and benchmark methods with 100 and 200 clients. For the Kermany dataset, we considered 5% participation, while for the RSNA dataset, we considered 10% participation. The image classification task involved training each algorithm for 2000 communication rounds. PFedHN was trained for 4000 global communication rounds to ensure equivalent communication costs. For the detection task, the methods were trained for 300 communication rounds.

Both tasks underwent optimization with an SGD optimizer, employing a default learning rate (lr) of 0.01 and a

	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. The TransFed method's average test accuracy is computed alongside that of multiple transformer-based approaches, encompassing different non-IID scenarios.

batch size (B) of 32. For TransFed, the Learnable generators were optimized using an SGD optimizer with a default learning rate (β) set to 0.01. The experiments were conducted on a cluster equipped with an NVIDIA Tesla V100 GPU, where the server and all clients were simulated.

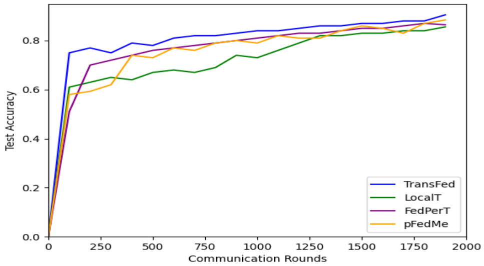

Figure 3. Test accuracy and convergence behavior of TransFed and other transformer-based methods on RSNA dataset.

4.2. Model Evaluation

We assessed model performance at 5-round intervals over the last 300 global communication rounds. Mean accuracy and standard deviation were computed. The average accuracy for each evaluation step was determined by the ratio of correct predictions to the total number of test images.

4.2.1 Performance Analysis

We conducted a comprehensive performance comparison between TransFed and several well-known federated learning methods, designed initially based on CNN backbones. Table 2 displays the average test accuracy of these algorithms, highlighting TransFed's remarkable performance superiority over each of them. This result strongly supports the assertions made in our Introduction section: 1) The FedAvg algorithm could impede the distinctive client representations within transformer models, as evidenced by Local-T outperforming FedAvg-T; and 2) TransFed's learned customized focal modulation effectively addresses data heterogeneity. As depicted in Table 2, TransFed consistently outperforms Vanilla customized-T across all settings, thereby confirming that the "learn-to-tailor" approach leverages the strengths of focal-modulation in transformer models.

To deepen our understanding, we conducted an evaluation of test accuracy along with the curve depicting global communication rounds in TransFed. This comparison was extended to other transformer-based methods, as depicted in Figure 3. The analysis reveals that TransFed demonstrates a smooth curve and achieves higher accuracy compared to alternative approaches.

4.2.2 Analysis of Different Adapted Parts

This study examined the effects of personalizing various components of the transformer model. Specifically, we focused on four components: (1) the focal-modulation layers (our proposed method), (2) the Multi layer perceptron layers, (3) the normalization layers, and (4) the encoder. To maintain a fair comparison, we employed the identical Learnable generator to generate the parameters associated with these individual components while ensuring consistency in the focalnet structures as described. The results of this experiment are presented in Table 3. It is evident from the table that personalizing the focal-modulation layers yields the best performance compared to personalizing other components. Furthermore, Table 3 illustrates that customizing the normalization layers yields superior performance compared to tailoring the MLP layers and the absolute encoder.

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.

4.2.3 Generalization to Novel Clients

We thoroughly assessed our method's capacity for generalization, contrasting it with pFedMe, pFedHN, FedRod, and a customized-T Vanilla approach on the Kermany and

RSNA datasets under the Beta configuration. To simulate a realistic scenario, 30% of the clients were randomly selected as novel clients whose data had not been seen during the training phase. FedPer fine-tuned the customized parameters in the last classification layer, while pFedMe learned all parameters to obtain customized models for each client. In the case of pFedHN and TransFed, the customized parameters had the option to choose between clients embedding vectors with a dimension of 32 and the entire Learnable generator. As presented in Table 4, the outcomes suggest that TransFed (Learnable generator) exhibits improved resilience and adeptly adjusts to new clients in few epochs.

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.

4.2.4 Analysis of Learnable generators

To thoroughly examine the impact of Learnable generators, we conducted a comparative analysis between TransFed and Vanilla customized-T. The latter method restores the projection parameters P_i for each client locally, without the utilization of Learnable generators. As depicted in Table 2, TransFed exhibits a significant advantage over Vanilla customized-T, highlighting the crucial role of Learnable generators in TransFed. We also observed that even when Learnable generators solely generate the parameters of the focal-modulation layer, they effectively encode client-specific information into client embeddings z_i. The Learnable generators have the capability to project client embeddings z_i onto a manifold defined by the parameters ϕ of the Learnable generator.

To delve deeper into the acquired client embeddings, we utilized the t-SNE algorithm for their projection onto a 2-D plane, as illustrated in Figure 4. Specifically, we distributed each coarse class among five clients, ensuring that the corresponding fine classes were uniformly allocated among these selected clients. We also trained TransFed and visualized the client embeddings after training. Learned individual embeddings of clients who share common coarse labels tend to cluster together, while those with dissimilar coarse labels are mapped farther apart. This outcome provides compelling evidence supporting our claim that Learnable generators are highly effective for encoding customized information into client embeddings z_i.

4.3. Ablation Study

Data heterogeneity, especially in label distribution, is a significant challenge in customized federated learning. We conducted experiments on RSNA and Kermany datasets, varying the parameter alpha of the Beta distribution. Smaller alpha values indicate a higher level of heterogeneity. TransFed consistently outperformed benchmark methods (FedAvg-T, FedBN [16], and pFedHN [9]) in handling label distribution heterogeneity. It demonstrated robustness even when other methods struggled to utilize client heterogeneity effectively. We also explored the impact of noise-induced feature imbalance on TransFed. By adding Gaussian noise with increasing levels to each client, we assessed TransFed's performance. It consistently outperformed other methods in handling client-specific noise. The number of focal-modulation blocks in TransFed was investigated, showing that increasing the number improved the model's ability to capture data heterogeneity and enhance overall performance (Table 3). Consequently, we selected eight as the default attention block number for TransFed in subsequent experiments. We examined the impact of the number of participating clients on model performance by varying the sample rate. TransFed exhibited greater stability than FedAvg-T, as shown in Figure 3. This highlights the robustness of TransFed in handling different client participation rates.

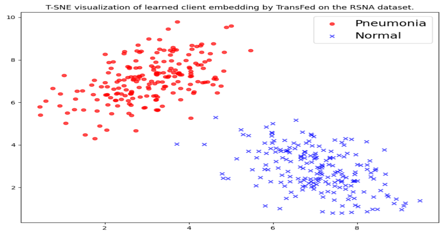

Figure 4. **Visualization of Client Embeddings Learned by TransFed using t-SNE on the RSNA Dataset.**

5. Conclusion and Future Work

We introduced TransFed, a transformer-based federated learning framework that addresses the limitations of Focal Modulation in non-IID scenarios. TransFed overcomes the degradation of Focal Modulation under traditional Federated Averaging (FedAvg) by adopting customized Focal Modulation for each client. TransFed enhances the performance of Focal Modulation by tailoring it to each client through the use of a central Learnable generator. This collaborative approach improves scalability and generalization. Experimental results demonstrate TransFed's superiority in non-IID contexts, showcasing its effectiveness against noise in local datasets. TransFed's potential synergies with advanced federated methods are promising. While TransFed's training speed lags behind CNN models, ongoing research aims to enhance its computational and communication efficiency.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 2

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 3, 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification, 2022. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020. 3

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[9] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3, 5, 8

[10] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7865–7873, 2021. 3

[11] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2):651, 2018. 6

[12] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. 3

[13] 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. 3, 6

[14] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021. 3

[15] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 3, 6

[16] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021. 8

[17] Bingyan Liu, Yao Guo, and Xiangqun Chen. Pfa: Privacy-preserving federated adaptation for effective model personalization. In *Proceedings of the Web Conference 2021*, pages 923–934, 2021. 3

[18] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10092–10101, 2022. 3

[19] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020. 3

[20] Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, pages 15070–15092. PMLR, 2022. 3

[21] 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

[22] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 3

[23] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10061–10071, 2022. 3

[24] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 1

[25] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020. 1, 3

[26] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019. 1

[27] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020. 3, 6

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[29] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019. 3

[30] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 2

[31] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2, 6

[32] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 1, 2, 3

[33] Leijie Zhang, Ye Shi, Yu-Cheng Chang, and Chin-Teng Lin. Federated fuzzy neural network with evolutionary rule learning. *IEEE Transactions on Fuzzy Systems*, 2022. 3

[34] Weishan Zhang, Xiao Chen, Ke He, Leiming Chen, Liang Xu, Xiao Wang, and Su Yang. Semi-asynchronous personalized federated learning for short-term photovoltaic power forecasting. *Digital Communications and Networks*, 2022. 3

[35] Weishan Zhang, Fa Yu, Xiao Wang, Xingjie Zeng, Hongwei Zhao, Yonglin Tian, Fei-Yue Wang, Longfei Li, and Zengxiang Li. R^{2} fed: Resilient reinforcement federated learning for industrial applications. *IEEE Transactions on Industrial Informatics*, 2022. 3

[36] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pages 12878–12889. PMLR, 2021. 3

[37] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 2