Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning

Liangqiong Qu[1*] Yuyin Zhou[2*] Paul Pu Liang[3*] Yingda Xia[4] Feifei Wang[1]
Ehsan Adeli[1] Li Fei-Fei[1] Daniel Rubin[1]

[1] Stanford University,[2] UC Santa Cruz, [3] Carnegie Mellon University, [4] Johns Hopkins University

{liangqiqu, zhouyuyiner, philyingdaxia}@gmail.com, pliang@cs.cmu.edu,
{ffwang, eadeli, feifeili, rubin}@stanford.edu,

Abstract

*Federated learning is an emerging research paradigm enabling collaborative training of machine learning models among different organizations while keeping data private at each institution. Despite recent progress, there remain fundamental challenges such as the lack of convergence and the potential for catastrophic forgetting across real-world heterogeneous devices. In this paper, we demonstrate that self-attention-based architectures (*e.g.*, Transformers) are more robust to distribution shifts and hence improve federated learning over heterogeneous data. Concretely, we conduct the first rigorous empirical investigation of different neural architectures across a range of federated algorithms, real-world benchmarks, and heterogeneous data splits. Our experiments show that simply replacing convolutional networks with Transformers can greatly reduce catastrophic forgetting of previous devices, accelerate convergence, and reach a better global model, especially when dealing with heterogeneous data. We release our code and pretrained models to encourage future exploration in robust architectures as an alternative to current research efforts on the optimization front.*

1. Introduction

Federated Learning (FL) is an emerging research paradigm to train machine learning models on private data distributed over multiple heterogeneous devices [47]. FL keeps data on each device private and aims to train a global model that is updated only via communicated parameters instead of the data itself. Therefore, it provides an opportunity for collaborative machine learning across multiple institutions without risking leakage of private data [25,36,54]. This has proved especially useful in domains such as healthcare [4, 7, 15, 40], learning from mobile devices [17, 38], smart cities [25], and communication networks [49], where preserving privacy is crucial. Despite the rich opportunities afforded by FL, there remain fundamental research problems to be tackled before FL can be readily applicable to real-world data distributions. Most current methods that aim to learn a single global model across non-IID devices encounter challenges such as non-guaranteed convergence and model weight divergence for parallel FL meth-

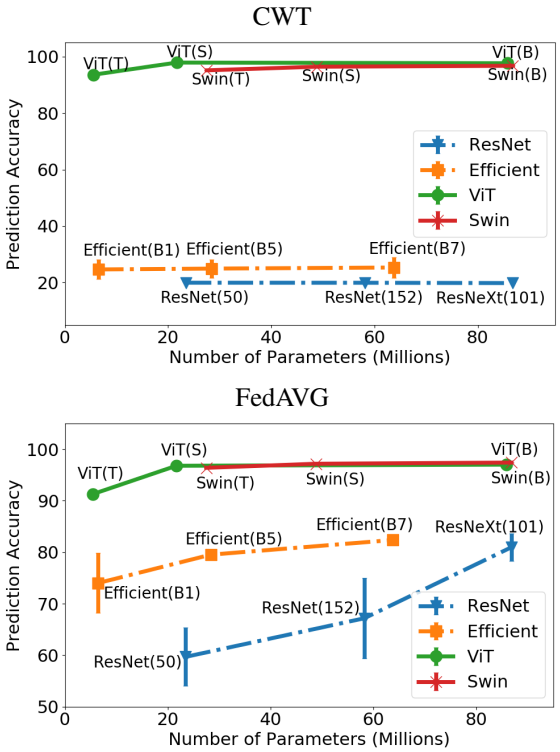

Figure 1. Prediction test accuracy on highly heterogeneous data partitions (Split-3) of CIFAR-10 dataset versus model size[1]. Vision Transformers (ViTs and Swin Transformers) significantly outperform CNNs (ResNets and EfficientNets) on highly heterogeneous data partitions.

*Equal contribution

ods [35,37,68], and severe catastrophic forgetting problems for serial FL methods [7, 16, 57].

While most research efforts focus on improving the optimization process in FL, our paper aims to provide a new perspective by rethinking the choice of architectures in federated models. We hypothesize that Transformer architectures [12, 61] are especially suitable for heterogeneous data distributions due to their surprising robustness to distribution shifts [3]. This property has led to the prevalence of Transformers in self-supervised learning where heterogeneity is manifested via distribution shifts between unlabeled pretraining data and labeled test data [11], as well as in multimodal learning over fundamentally heterogeneous input modalities such as image and text [24, 60]. To study this hypothesis, we conduct the first large-scale empirical benchmarking of several neural architectures across a suite of federated algorithms, real-world benchmarks, and heterogeneous data splits. To represent Transformer networks, we use a standard implementation of Vision Transformers [12, 41] on image tasks spanning image classification [31, 42] and medical image classification [27].

Our results suggest that ViT-FL (Federated Learning with Vision Transformers) performs especially well in settings with most heterogeneous device splits, with the gap between ViT-FL and FL with ResNets [19] increasing significantly as heterogeneity increases. To understand these results, we find that the main source of improvement lies in the increased robustness of Transformer models to heterogeneous data which reduces catastrophic forgetting of previous devices when trained on substantially different new ones. Together, Transformers converge faster and reach a better global model suitable for most devices. Through comparisons to FL methods designed specifically to combat heterogeneous data, we find that ViT-FL provides immediate improvements without using training heuristics, additional hyperparameter tuning, or additional training. Moreover, it is noteworthy that our ViT-FL is orthogonal to existing optimization based FL methods, and can be easily applied to improve their performance. To this end, we conclude that Transformers should be regarded as a natural starting point for FL problems in future research.

2. Related Work

Federated Learning. Federated learning (FL) aims to train machine learning models on private data across massively distributed devices [47]. To enable effective distributed training across heterogeneous devices, two categories of methods have emerged: (1) parallel FL methods involve training each local client in parallel either synchronously or asynchronously (such as the classic FedAVG [47]), whereas (2) serial methods train each client in a serial and cyclical way (such as Cyclic Weight Transfer (CWT) [7] and Split learning [62]). A schematic description of FedAVG [47] and CWT [7] is illustrated in Figure 2. At its core, FL presents a challenge of data heterogeneity in the distributions of training data across clients, which causes non-guaranteed convergence and model weight divergence for parallel FL methods [21, 37, 66, 68], and severe catastrophic forgetting problem for serial FL methods [7, 16, 57].

Among recent developments to the classic FedAVG algorithm [47] have included using server momentum (FedAVGM) to mitigate per-client distribution shift and imbalance [22], globally sharing small subsets of data among all users (FedAVG-Share) [68], using a proximal term to the local objective (FedProx) to reduce potential weight divergence [37], or using other optimization heuristics such as collaborative replay [52], unsupervised contrastive learning [69], matching feature layers of user models [64, 65], or model distillation [14] to handle heterogeneity.

Concurrently, several recent efforts aim to alleviate catastrophic forgetting in continual and serial learning: constraining the updates on weights that are important to previously seen tasks or clients (elastic weight consolidation (EWC) [30]), applying Deep Generative Replay to mimic data from previous clients or tasks [52, 58], and applying cyclically weighted objectives to mitigate performance loss across label distribution skewness [2], among others. However, all of these approaches mainly focus on improving the optimization algorithm without studying the potential in architecture design to improve robustness to distribution shifts in data. In our work, we show that simple choices in architecture actually make a big difference and should be an active area of study in parallel to the optimization methods that have been the main focus of current work.

Transformers. The Transformer architecture was first proposed for sequence-to-sequence machine translation [61] and has subsequently established state-of-the-art performance across many NLP tasks, especially when trained in a self-supervised paradigm [11]. Recently, Transformers have also been found to be broadly applicable to tasks involving images and video. For instance, Parmar *et al.* [50] applied self-attention to local neighborhoods of an image while the Vision Transformer (ViT) [12] achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images.

Its intriguing performance boosts relative to classical architectures for language (*i.e.*, LSTMs [20]) and vision (*i.e.*, CNNs [19, 34]) have inspired recent interest towards understanding the reasons behind their effectiveness. Among several particularly relevant findings are that ViTs are highly robust to severe occlusions, perturbations, domain shifts [3, 48], as well as synthetic and natural adversarial examples [44, 51]. In addition, recent studies have suggested that Transformers are also suitable for heterogeneous and multimodal data [24, 43, 60]. Inspired by these findings, we hy-

[1] Mean and standard deviation are calculated across three runs.

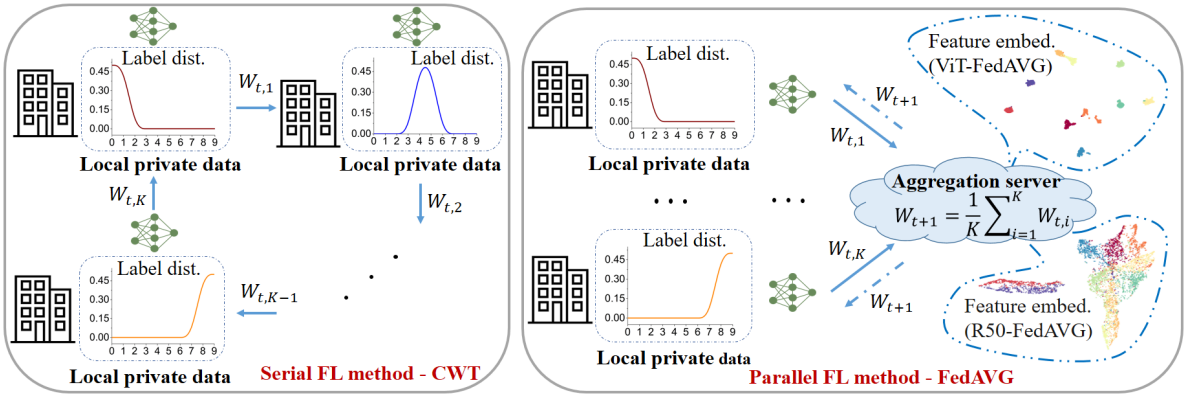

Figure 2. Simplified schematic for a typical serial FL method CWT [7] and a parallel FL method FedAVG [46] on non-IID data partitions of CIFAR-10 [31] with label distribution skewness. $W_{t,i}$ denotes the model weights during training at round t on client i (total K clients are involved). On the right, we show feature embedding visualizations of ViT(S)-FedAVG and ResNet(50)-FedAVG using UMAP [45]. We find that the features learned by ViT(S)-FedAVG are more clearly separated than those learned by ResNet(50)-FedAVG. Our experiments (section 4.2) support the superiority of ViT-FL on heterogeneous data and we provide analysis explaining their effectiveness (section 4.3).

pothesize that ViTs will also be highly effective in adapting to data heterogeneity in FL, and provide detailed empirical analysis to test this hypothesis.

3. Transformers in Federated Learning

In this section, we present background on Transformer architectures and federated learning methods.

3.1. Vision Architectures

CNN. For convolution-based architectures, we use the ResNet [19] model family (ResNet-50, ResNet-152, and ResNeXt-101 (32x8d)) and EfficientNet [59] model family (EfficientNet-B1, EfficientNet-B5, and EfficientNet-B7), which contains a sequence of convolution, ReLU, pooling, and batch normalization layers. ResNet and EfficientNet are among the most popular architectures for image classification and have been the standard architecture used in FL on image data [1, 39].

Transformers. As a comparison, we employ Vision Transformers (ViT(S), ViT(T), ViT(B)) [12] model family and Swin Transformer model family (Swin(T), Swin(S), and Swin(B)) [41], which do not use conventional convolution layers. Instead, the image features are extracted with image sequentialization and patch embedding strategies. See Figure 1 for the number of parameters for each model.

3.2. Federated Learning Methods

We apply one of the most popular parallel methods (FedAVG [47]) and serial methods (CWT [7]) as training algorithms (see schematic descriptions in Figure 2).

Federated Averaging. FedAVG combines local stochastic gradient descent (SGD) on each client with iterative model averaging [47]. Specifically, a fraction of local clients are randomly sampled in each communication round, and the server sends the current global model to each of these clients. Each selected client then performs E epochs of local SGD on its local training data and sends the local gradients back to the central server for aggregation synchronously. The server then applies the averaged gradients to update its global model, and the process repeats.

Cyclic Weight Transfer. In contrast to FedAVG where each local client is trained in a synchronous and parallel way, the local clients in CWT are trained in a serial and cyclic manner. In each round of training, CWT trains a global model on one local client with its local data for a number of epochs E, and then transfers this global model to the next client for training, until all local clients have been trained on once [7]. The training process then cycles through the clients repeatedly until the model converges or a predefined number of communication rounds is reached.

4. Experiments

Our experiments are designed to answer the following research questions that are of importance to practical deployment of FL methods, while also aiding our understanding of (vision) Transformer architectures.

- Are Transformers able to learn a better global model in FL settings as compared to CNNs which have been the de-facto approach on FL tasks (section 4.2)?
- Are Transformers especially capable of handling heterogeneous data partitions (section 4.3.1)?
- Do Transformers reduce communication costs as compared to CNNs (section 4.3.2)?
- Can Transformers be applied to further improve existing optimization-based FL methods (section 4.4)?
- What are practical tips helpful for practitioners to deploy Transformers in FL (section 4.5)?

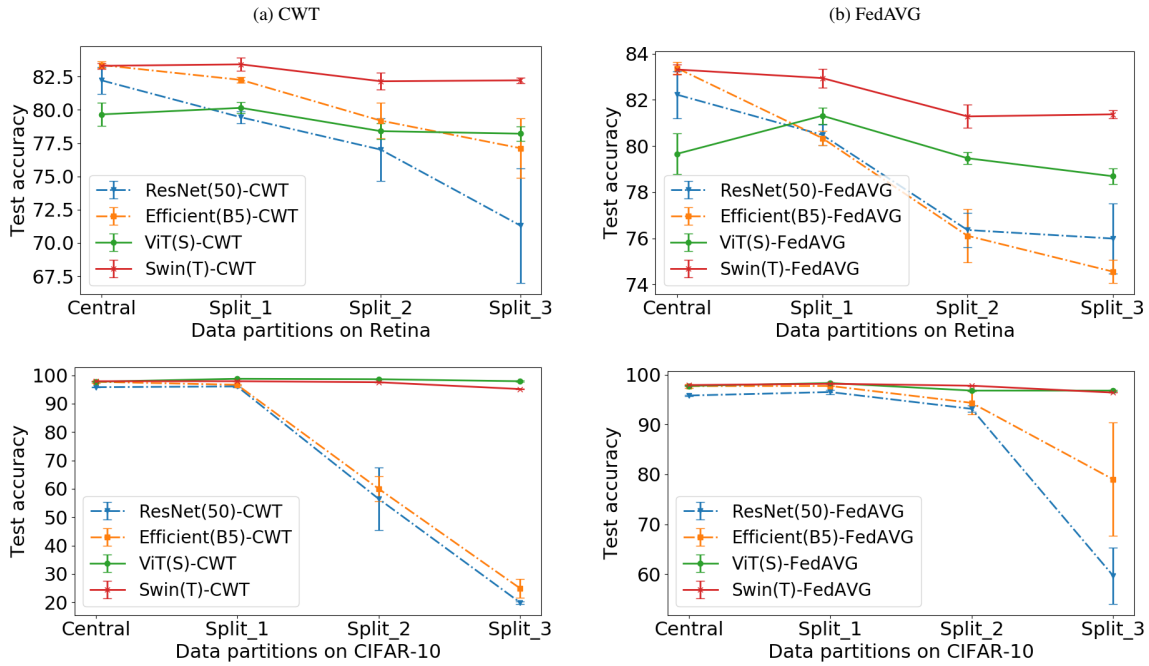

Figure 3. Prediction accuracy (%) of both CWT and FedAVG with CNNs and Transformers as baseline networks on Retina dataset (first row) and CIFAR-10 dataset (second row), respectively. Vision Transformers (both ViT and Swin) show consistently strong performance especially in non-IID data partitions.

Experimental code is included https://github.com/Liangqiong/ViT-FL-main .

4.1. Experimental Setup

Following [7,21], we evaluate FL on the Kaggle Diabetic Retinopathy competition dataset (denoted as Retina) [27], CIFAR-10 dataset [31] with simulated data partitions, and a real-world CelebA dataset [42] in our study.

Retina and CIFAR-10: We binarize the labels in the Retina dataset to Healthy (positive) and Diseased (negative), randomly selecting $6,000$ balanced images for training, $3,000$ images as the global validation dataset, and $3,000$ images as the global testing dataset following [7]. We use the original test set in CIFAR-10 as the global test dataset, set aside $5,000$ images from the original training dataset as the global validation dataset, and use the remaining $45,000$ images as the training dataset. We simulate three sets of data partitions: one IID-data partition, and two non-IID data partitions with label distribution skew. Each data partition in Retina and CIFAR-10 contains 4 and 5 simulated clients, respectively. We use the mean Kolmogorov-Smirnov (KS) statistic between every two clients to measure the degree of label distribution skewness. KS $= 0$ indicates IID data partitions, while KS $= 1$ results in an extremely non-IID data partition, where each client holds totally different label distributions (see Appendix A.1 for detailed pre-processing and data partitions).

CelebA is a large-scale face attributes dataset with more than 200K celebrity images. We use the federated version of CelebA provided by the LEAF benchmark [5] which partitions into devices based on identity. Following [5], we test on the binary classification task (presence of smile) and drop clients with larger than 8 samples to increase task difficulty. This results in a total of 227 clients each with an average of 5.34 ± 1.11 samples and a total of 1213 samples.

We use linear learning rate warm-up and decay scheduler for ViT-FL. The learning rate scheduler for FL with CNNs is selected from linear warm-up and decay or step decay. Gradient clipping (at global norm 1) is applied to stabilize training. We set the local training epoch E in all FL methods to 1 (unless otherwise stated), and the total communication rounds to 100 for Retina and CIFAR-10, and 30 for CelebA. For fair comparison, all models used in this paper are pretrained on ImageNet-1K [10]. More implementation details are in Appendix A.2.

Compute: All experiments were conducted on either a TITAN V GPU or Tesla V100 GPU.

4.2. Results

Comparison of FL with different neural architectures and (ideal) centralized training: Both CWT and FedAVG achieve comparable results to a model trained on centrally hosted data (denoted as Central) on the IID setting no matter which architecture is applied (Figure 3). However, we observe a significant reduction in test accuracy for CNNs on heterogeneous data partitions for both CWT and FedAVG,

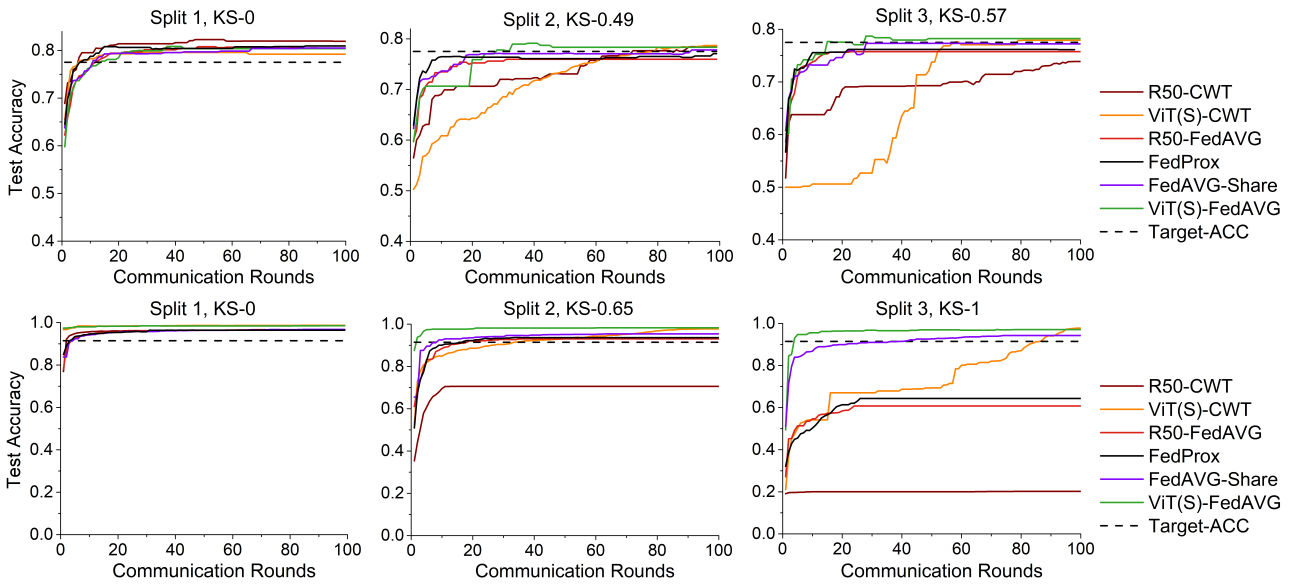

Figure 4. Test set accuracy versus communication rounds on Retina dataset (first row) and CIFAR-10 dataset (second row) with different data partitions. The black dashed line shows the target performance (Target-ACC) used in Table 3. Vision Transformers converge faster with fewer communication rounds, which make them especially suitable for communication-efficient FL.

especially on extremely heterogeneous data partitions (Split 3, KS-1 of CIFAR-10) (Figure 3 and Figure 1). By simply replacing CNNs with ViTs, both CWT and FedAVG successfully retain model accuracy even in highly heterogeneous non-IID settings. ViT(S)-CWT and ViT(S)-FedAVG improve the test accuracy relative to ResNet(50)-CWT and ResNet(50)-FedAVG by 77.70% and 37.34% on the highly heterogeneous Split-3, KS-1 of CIFAR-10 dataset. Therefore, ViT is particularly suitable for heterogeneous data.

Comparison with existing FL methods: We also compare ViT-FL to two state-of-the-art optimization based FL methods: FedProx [37], and FedAVG-Share [68] on both Retina and CIFAR-10. We use ResNet(50) as the backbone network for the other compared methods, and ViT(S) for our methods. We tune the best parameters (penalty constant μ in the proximal term of FedProx) on Split-2 dataset with grid search, and apply the same parameters to all the remaining data partitions. We allow each client to share 5% percentage of their data among each other for FedAVG-Share. As shown in Figure 4, ViT-FL outperforms all the other FL methods in non-IID data partitions, especially on the highly heterogeneous non-IID settings. FedProx [37] suffers severe performance drops on highly heterogeneous data partitions despite carefully tuned optimization parameters. Similarly, FedAVG-Share also suffers from performance drops on highly heterogeneous data partition Split-3 even when 5% percentage of the local data is shared among all clients (94.4% of Split-3 on CIFAR-10 dataset compared to 97% on Split-1). We conclude that simply using Transformers outperforms several recent methods designed for FL, which often require careful tuning of optimization parameters. Please note that the usage of ViTs is orthogonal to the existing optimization methods, and a combination of both can yield stronger performance (see details in Section 4.4).

4.3. Analyzing the Effectiveness of Transformers

Given these promising empirical results, we now perform a careful empirical analysis to uncover what exactly contributes to Transformers' improved performance.

4.3.1 Transformers generalize better in non-IID settings

The distributed nature of FL means that there can be substantial heterogeneity in data distributions across clients. Prior research has shown that training FL models with FedAVG or CWT incurs issues such as weight divergence and catastrophic forgetting respectively [30, 57]. We argue that the local convolutions used in CNNs, which have been shown to rely more on local high-frequency patterns [13, 26, 63], might be particularly sensitive to heterogeneous devices. This problem is particularly prevalent in FL over healthcare data since input images captured by different institutions may vary significantly in local patterns (intensity, contrast, *etc.*) due to different medical imaging protocols [16,55], as well as in natural data splits due to user idiosyncrasies in speaking [33], typing [17], and writing [28]. On the other hand, ViTs use self-attention to learn global interactions [53] and have been shown to be less biased towards local patterns as compared to CNNs. This property may contribute to their surprising robustness to distribution

	R50-CWT	ViT(S)-CWT	R50-FedAVG	R50-FedProx	R50-FedAVG-Share	ViT(S)-FedAVG
CelebA	85.35 ± 8.27	$\mathbf{88.09 \pm 5.15}$	84.08 ± 9.65	84.27 ± 9.74	85.46 ± 3.75	$\mathbf{86.63 \pm 7.12}$

Table 1. Prediction accuracy (%) on CelebA dataset. Vision Transformers show superior performance to their ResNet(50) (R50 in Table) counterparts, and also outperform the optimization based FL methods (FedProx and FedAVG-Share) with ResNet(50) as backbone network.

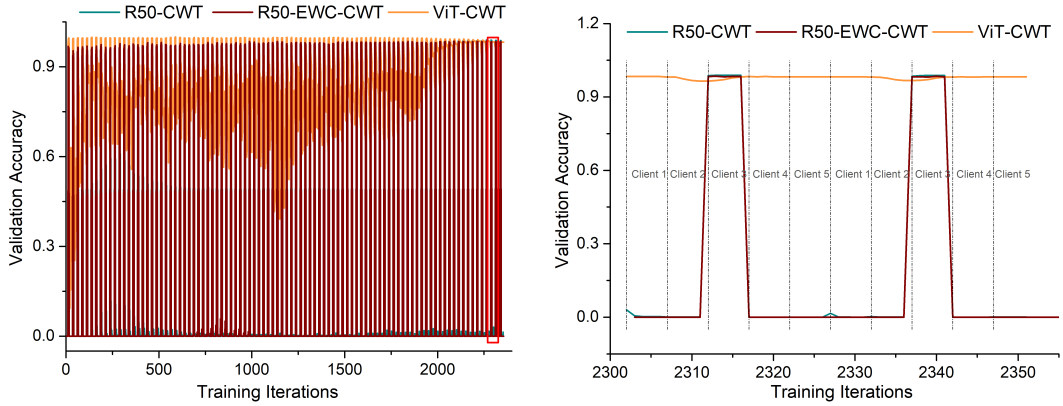

Figure 5. Left: evolution of the prediction accuracy on the validation dataset of client 3 as more clients are involved in CWT learning. We use Split 3 of CIFAR-10 dataset (most heterogeneous data split) and compare CWT trained with the ResNet(50) (R50 in Figure), ResNet(50)-EWC [30], and ViT(S) models. Right: zoom in on the red rectangular in the left image. The training order of different clients is also shown. The sequential training strategy of ResNet(50)-CWT incurs catastrophic forgetting on previous clients under highly heterogeneous data distributions. ResNet(50)-EWC-CWT [30] barely solves the catastrophic forgetting problem. ViT(S)-CWT helps alleviate this problem due to its strong generalization ability and robustness to heterogeneous data.

	RETINA (#6,000)		CIFAR-10 (#45,000)	
	CWT	FedAVG	CWT	FedAVG
ResNet(50)	51.3 ± 1.3	55.0 ± 0.3	31.2 ± 12.2	37.5 ± 1.4
ViT(S)	80.0 ± 0.1	81.0 ± 0.1	97.5 ± 0.02	97.4 ± 0.03

Table 2. Prediction accuracy (%) on a large-scale edge case setting with thousands of clients involved in training ($6,000$ and $45,000$ clients for Retina and CIFAR-10 respectively, with each client containing one data sample). Vision Transformers significantly outperform their ResNet counterparts in this edge case setting.

shifts and adversarial perturbations [3, 48]. To further analyze the generalization capabilities of Transformers across heterogeneous data, we design the following experiments:

1. Catastrophic forgetting across heterogeneous devices: CNNs often work worse on out-of-distribution data. This phenomenon is especially severe in the serial FL method CWT. Due to its sequential and serial training strategy, training CNNs in a CWT paradigm usually results in catastrophic forgetting on non-IID data partitions: the model's performance on previous clients abruptly degrades after a few updates on a new client with a different data distribution [3, 48]. This results in poorer and slower convergence which is undesirable in FL. Similar forgetting issues have also been found in the transfer learning literature [8, 9, 56].

We evaluate CWT on Split-3 of the CIFAR-10 dataset to illustrate this catastrophic forgetting phenomenon. In Figure 5, we plot the evolution of the prediction accuracy on the validation dataset of Client-3 (which shares the same data distribution as its training dataset) as more clients are involved in CWT learning. When transferring a well-trained model on Client-3 to Client-4, the prediction accuracy on the previous Client-3 validation dataset degrades abruptly and dramatically (from $> 98\%$ to $< 1\%$ accuracy). However, the model trained with ViT as backbone (ViT(S)-CWT) is able to transfer knowledge from Client-3 to Client-4 while losing only small amounts of information on Client-3 (maintains accuracy at 98%). Therefore, ViTs generalize better to new data distributions without forgetting old ones.

We further compare ViT(S)-CWT with an optimization method specifically designed to alleviate catastrophic forgetting, EWC [30] (using the implementation from [23]). Serial training of CWT on Split-3 of CIFAR-10 can be considered as an incremental class learning task where each client contains an exclusive subset of classes in the dataset. Each client model shares the same classifier to a standardized union label space [23]. However, from Figure 5, EWC barely solves the catastrophic forgetting problem on highly heterogeneous data partitions, which also matches the results reported in [23]. This experiment further demonstrates the effectiveness of ViT beyond optimization methods designed for FL.

2. Generalization of VIT-FL on real-world federated datasets: A well-trained federated model should perform well on out-of-distribution test datasets of other unseen clients. To test the generalizability of Transformers, we

		CWT		FEDAVG					
		R50	ViT(S)	R50	R50-FedProx	R50-Share	ViT(S)	ViT(S)-FedProx	ViT(S)-Share
RETINA	Split-1	**6 × 23.5**	9 × 21.7	12 × 23.5	7 × 23.5	11 × 23.5	11 × 21.4	**4 × 21.4**	7 × 21.4
	Split-2	72 × 23.5	55 × 21.4	∞	∞	85	15 × 21.4	**12 × 21.4**	**13 × 21.4**
	Split-3	∞	58 × 21.4	∞	∞	∞	**15 × 21.4**	**12 × 21.4**	16 × 21.4
CIFAR-10	Split-1	2 × 23.5	**1 × 21.4**	4× 23.5	4 × 23.5	5 × 23.5	**1 × 21.4**	**1 × 21.4**	**1 × 21.4**
	Split-2	∞	34 × 21.7	19 × 23.5	17 × 23.5	9 × 23.5	**2 × 21.4**	**2 × 21.4**	**1 × 21.4**
	Split-3	∞	85 × 21.7	∞	∞	41 × 23.5	4 × 21.4	**3 × 21.4**	**1 × 21.4**

Table 3. # transmitted message size (# communication round × # model parameters (M)) required to reach target performance (**best** and **second best**). # model parameters of ViT(S) and ResNet(50) is 21.7M and 23.5M, respectively. ViTs converge faster especially on heterogeneous data splits, and can be combined with optimization-based methods (FedProx and FedAVG-Share) for even faster convergence.

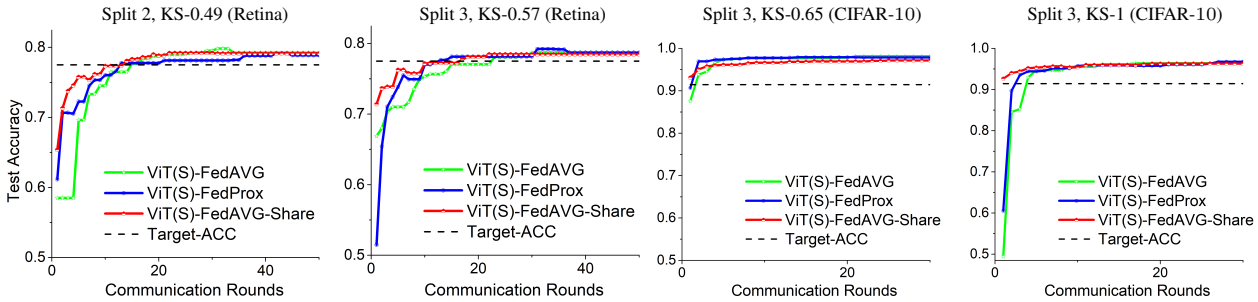

Figure 6. Test set accuracy versus communication rounds on ViT(S)-FedAVG and their combination with existing FL methods FedProx [37] and FedAVG-Share [68]. Vision Transformers can be used in conjunction with existing optimization based FL methods to further improve convergence speed and reach target performance with fewer communication rounds.

apply it to a real-world federated CelebA dataset [42] and compare it to the ResNet counterparts, FedProx [37], and FedAVG-Share [68]. We report the test accuracies of models trained using different FL methods on the union of the test data from all local clients in Table 1. Our VIT-FL approach outperforms state-of-the-art FL methods, and also reduces variance. This shows that Transformers learn a better global model than their CNNs counterparts.

3. Generalization of VIT-FL on extreme large-scale setting: To validate the effectiveness of VIT-FL on a more large-scale real-world distributed learning setting where thousands of clients are involved, we further apply different FL methods to an extreme edge case situation on both Retina and CIFAR-10 dataset. The edge case here is defined as one client holding *only one* data sample, which is quite common in healthcare where the patient holds only one data sample belonging to themselves. This results in an extremely large number of heterogeneous clients: $6,000$ for Retina and $45,000$ for CIFAR-10. From Table 2, ViTs still learn a promising global model on this extremely heterogeneous edge case setting, significantly outperforming ResNet models (from 50% to 80% on Retina and from 30% to 90% on CIFAR-10).

4.3.2 Transformers converge faster to better optimum

A powerful FL method should not only perform robustly on both IID and non-IID data partitions but also have low communication costs to enable deployment over communicated-limited bandwidths. Communication cost is determined by the number of rounds till convergence and the number of model parameter. We calculate the number of communication rounds needed to achieve a predefined target test set accuracy of 95% of the prediction accuracy of a centrally trained ResNet(50). Specifically, we set the target accuracy of Retina and CIFAR-10 dataset to be 77.5% and 91.5% respectively. We define one communication round on the serial CWT method as one complete training cycle across all federated local clients.

From Figure 4 and Table 3, all the evaluated FL methods converge to the target test performance quickly on homogeneous data partitions. However, the convergence speed of ResNet(50)-FedAVG and ResNet(50)-CWT decrease with increasing heterogeneity and even reach a plateau on highly heterogeneous data partitions (and never reach the target accuracy). In contrast, VIT-FL still converges quickly on heterogeneous data. For example, ResNet(50)-CWT completely diverges due to severe catastrophic forgetting on heterogeneous data partitions Split-2 and Split-3 on CIFAR-10, whereas ViT(S)-CWT reaches the target performance after 34 and 85 communication rounds.

4.4. In Conjunction with Existing Methods

Since our investigation into architectural choices is largely orthogonal to existing optimization based FL methods, our findings can be easily used in conjunction with the latter. We combine Vision Transformers with optimization-based methods (FedProx [37] and FedAVG-Share [68]), and apply it to both Retina and CIFAR-10 datasets. From Ta-

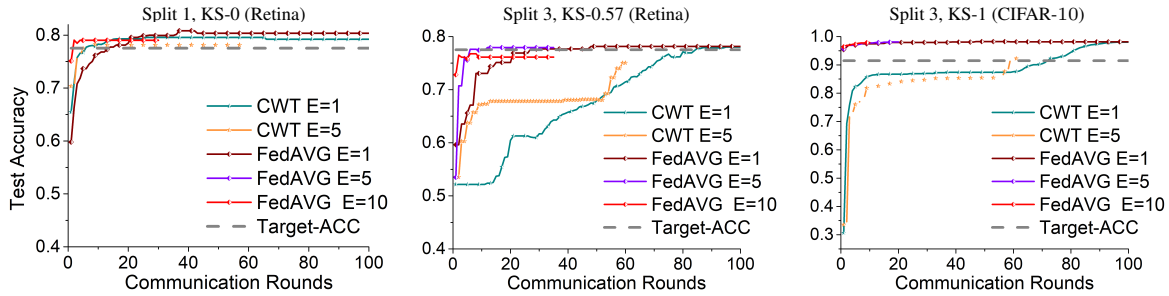

Figure 7. The effect of training over different local epochs E on each communication round for ViT(B)-CWT and ViT(B)-FedAVG models on Retina and CIFAR-10 (the ViT(B) prefix of CWT and FedAVG in the legend labels is omitted for simplicity). Large E leads to faster convergence in mild heterogeneous data partitions, but might lead to worse final performance in severely heterogeneous data partitions.

	CENTRAL	SPLIT-1	SPLIT-2	SPLIT-3
Pretrain	97.91	98.17	97.78	96.40
From scratch	94.50	86.91	79.43	64.50

Table 4. The influence of pretraining of Swin(T)-FedAVG on CIFAR-10. Similar to the training of VIT, pretraining is important for the training of VIT-FL.

ble 3 and Figure 6, when applying to existing FL optimization methods, VIT further boosts the performance for heterogeneous data clients.

4.5. Take-aways for Practical Usage

Local training epochs: It is standard to use E to denote the number of rounds a local model passes over its local dataset. E is known to strongly affect the performance of FedAVG [47] and CWT [7]. We conduct an experimental study on the impact of local training epochs E on VIT-FL. We consider $E \in \{1, 5, 10\}$ for ViT(B)-FedAVG, and $E \in \{1, 5\}$ for ViT(B)-CWT. From Figure 7, we find that ViT shows similar phenomena to their CNN counterparts, *i.e.*, larger E accelerates convergence of ViT(B)-FedAVG on homogeneous data partitions, but may lead to deterioration of final performance on heterogeneous data partitions. Similarly, ViT(B)-CWT also favors frequent transfer rate between each client as ResNet(50)-CWT [7] on non-IID data partitions. Therefore, we suggest users apply large E on homogeneous data to reduce communication, but a small E ($E \leq 5$ for VIT-FedAVG and $E = 1$ for VIT-CWT) for highly heterogeneous cases.

The influence of pretraining on VIT-FL: Evidence suggests that VIT generally require a larger amount of training data to perform better than CNNs when trained from scratch [12]. We conduct experiments to investigate the influence of pretraining on VIT-FL. We apply FedAVG as the training algorithm, use Swin(T) [41] as the backbone network, and test on CIFAR-10. We apply the same augmentation and regularization strategies as [41] during training and set the maximum communication rounds to 300. As shown in Table 4, the performance of Swin(T) drops when trained from scratch for both the ideal centrally-hosted and FL settings. Despite this, its performance on highly-heterogeneous data partition Split-3 when trained from scratch (64.50%) is surprisingly better than ResNet(50)-FedAVG (59.68% on Figure 3) when pretrained with orders of magnitude more data. In real applications, users are recommended to apply VIT as their first option, since VIT-FL consistently outperform their CNNs counterparts when pretrained models are applied (Figure 1 and Figure 3). If large-scale pretraining datasets are not available, self-supervised pretraining [6, 18] could be an alternative.

Other training tips: The training strategy of VIT in FL can be directly inherited from VIT training, such as using linear warm-up and learning rate decay, and gradient clipping. Relatively small learning rates and gradient norm clip are necessary to stabilize the training of VIT in CWT, especially in highly heterogeneous data partitions. Gradient norm clip also helps in the training of FL with CNNs across heterogeneous data since it has been shown to reduce weight divergence between local updates and the current global model [37]. Please refer to Appendix B.1 for more general tips and experimental analysis.

5. Conclusion

Despite the recent progress in FL, there remain challenges regarding convergence and forgetting when dealing with heterogeneous data. Unlike previous methods on improving optimization, we provide a new perspective by rethinking architecture design in FL. Using the robustness of Transformers to heterogeneous data and distribution shifts, we perform extensive analysis and demonstrate the advantages of Transformers in alleviating catastrophic forgetting, accelerating convergence, and reaching a better optimum for both parallel and serial FL methods. We release our code and models to encourage developments in robust architectures in parallel to efforts on the optimization front.

Acknowledgments

This work was supported in part by a grant from the NCI, U01CA242879.

References

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

[2] Niranjan Balachandar, Ken Chang, Jayashree Kalpathy-Cramer, and Daniel L Rubin. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *Journal of the American Medical Informatics Association*, 27(5):700–708, 2020. 2

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021. 2, 6

[4] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018. 1

[5] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018. 4, 12

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 8

[7] Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8):945–954, 2018. 1, 2, 3, 4, 8, 12

[8] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, 2019. 6

[9] Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 12

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] 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. In *ICLR*, 2021. 2, 3, 8, 12

[13] 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. *ICLR*, 2019. 5

[14] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15076–15086, October 2021. 2

[15] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M. Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2423–2432, June 2021. 1

[16] Sharut Gupta, Praveer Singh, Ken Chang, Liangqiong Qu, Mehak Aggarwal, Nishanth Arun, Ashwin Vaswani, Shruti Raghavan, Vibha Agarwal, Mishka Gidwani, et al. Addressing catastrophic forgetting for medical domain expansion. *arXiv preprint arXiv:2103.13511*, 2021. 2, 5, 13

[17] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. 1, 5

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 8

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 12, 14

[20] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[21] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020. 2, 4, 12, 13

[22] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 2, 13, 14

[23] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. In *NeurIPS Continual learning Workshop*, 2018. 6

[24] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021. 2

[25] Ji Chu Jiang, Burak Kantarci, Sema Oktug, and Tolga Soyata. Federated learning in smart city sensing: Challenges and opportunities. *Sensors*, 20(21):6230, 2020. 1

[26] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017. 5

[27] Kaggle. Diabetic retinopathy detection. *https://www.kaggle.com/c/diabetic-retinopathy-detection*, 2017. 2, 4, 12

[28] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 5

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12

[30] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 5, 6

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 3, 4, 12

[32] Fan Lai, Yinwei Dai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Fedscale: Benchmarking model and system performance of federated learning. In *Proceedings of the First Workshop on Systems Challenges in Reliable and Secure Federated Learning*, pages 1–3, 2021. 13

[33] Siddique Latif, Sara Khalifa, Rajib Rana, and Raja Jurdak. Federated learning for speech emotion recognition applications. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 341–342. IEEE, 2020. 5

[34] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[35] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10713–10722, June 2021. 2

[36] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. 1

[37] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018. 2, 5, 7, 8, 13, 14

[38] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 1

[39] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *arXiv preprint arXiv:2006.07242*, 2020. 3

[40] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1013–1023, June 2021. 1

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 3, 8, 14

[42] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 4, 7

[43] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021. 2

[44] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. *arXiv preprint arXiv:2104.02610*, 2021. 2

[45] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[46] 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. 3

[47] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016. 1, 2, 3, 8

[48] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021. 2, 6

[49] Solmaz Niknam, Harpreet S Dhillon, and Jeffrey H Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6):46–51, 2020. 1

[50] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 2

[51] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 2

[52] Liangqiong Qu, Niranjan Balachandar, Miao Zhang, and Daniel Rubin. Handling data heterogeneity with generative replay in collaborative learning for medical imaging. *arXiv preprint arXiv:2106.13208*, 2021. 2

[53] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *NeurIPS*, 2019. 5

[54] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020. 1

[55] Holger R Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C Bizzo, et al. Federated learning

for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 181–191. Springer, 2020. 5

[56] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 6

[57] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020. 2, 5

[58] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017. 2

[59] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3, 14

[60] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019. 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[62] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018. 2

[63] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 5

[64] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020. 2

[65] Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. Federated learning for non-iid data via unified feature learning and optimization objective alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4420–4428, October 2021. 2

[66] Miao Zhang, Liangqiong Qu, Praveer Singh, Jayashree Kalpathy-Cramer, and Daniel L Rubin. Splitavg: A heterogeneity-aware federated deep learning method for medical imaging. *arXiv preprint arXiv:2107.02375*, 2021. 2

[67] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *ICLR*, 2021. 13, 14

[68] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 2, 5, 7

[69] Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4912–4921, October 2021. 2