Performance Analysis of Vision Transformers with Federated Learning for Edge Devices

Jean Charle Yaacoub
j.yaacoub@mail.utoronto.ca

Adrian Yung
ad.yung@mail.utoronto.ca

Abstract

Vision Transformers (ViTs) have demonstrated comparable or better image classification performance than prior models while using comparable or less computational resources. Additionally, Federated Learning (FL) has emerged as the solution to private distributed learning by allowing for distributed training on separate client datasets. This paper investigates the performance of ViTs under the FL setting. The efficiency and accuracy of ViTs are compared to the performance of traditional Convolutional Neural Networks (CNNs) with ResNet-50 and modern-day CNNs like ConvNeXt. Results compare IID and non-IID settings under simulated and real environments, revealing that the original ViT model outperforms prior ResNet-50 models and is competitive with ConvNeXt in terms of accuracy. These findings highlight the value of investigating and optimizing models in a federated environment on edge devices and provide insight into the use of ViTs for FL with their efficiency and overall performance. The code is publicly available at: https://github.com/jyaacoub/FL-ViT.

1 Introduction

Visual media is an increasingly important source of information in today's digital age and can provide valuable insights into a wide range of applications including security, healthcare, and entertainment. Visual media analysis involves developing algorithms and methods capable of extracting useful information from visual media. Convolutional neural networks (CNNs) are one such method that has seen widespread usage in visual media analysis. Being quite effective at identifying spatial relationships in both text and image data, the CNN architecture builds up on local patches of input data. CNNs extract local features from the input images via convolutional layers, then combine them to create complex representations [1]. Recently, transformers have seen a surge in usage based on the mechanism of self-attention, prioritizing the most relevant areas in the input. Attention scores are calculated for each element in a sequence or set, then used to weight the importance of each element's representation. The model can then determine the relevant parts of the input and downplay the importance of irrelevant or noisy information [2]. Models with self-attention have become more prevalent in natural language processing (NLP) and computer vision tasks with a better ability to capture long-range dependencies and improve the quality of predictions [2]. Vision transformers (ViTs) introduced patch embeddings for the tokenization of images that bridged the gap between NLP and CV. ViTs, based on the Transformer model, have achieved state-of-the-art results on many image classification tasks with up to 4x the performance of CNNs in terms of computational efficiency and accuracy [3].

Traditional machine learning approaches largely depend on centralized data and processing, which can be problematic in situations with sensitive data such as healthcare data or financial information. In addition, centralized processing and data transfer can be computationally expensive and subject to regulations. Federated learning (FL) addresses these challenges by enabling models to be trained locally on data spread across multiple devices, with only model updates shared with a centralized server [4, 5]. This approach increases privacy and adheres to

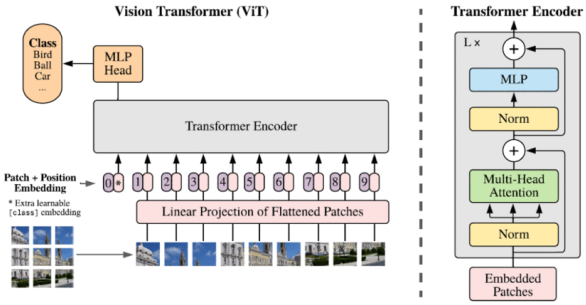

Figure 1: Architecture of the Vision Transformer [3]

regulations that prevent sharing of sensitive data. With the rapid advancements of new technology in the image recognition and classification space, it is intriguing to look into these topics, particularly ViTs and FL, more closely.

2 Background

The original ViT architecture adopted the transformer architecture with minimal changes to make it suitable for image-based tasks (Figure 1). ViT first divides an input image into a sequence of non-overlapping patches. The size and number of these patches depend on the ViT architecture and the size of the input image. Each patch is then flattened into a vector and projected into a lower-dimensional embedding space which are concatenated to form a sequence of embeddings, similar to a series of word embeddings used with text data. The sequence of embeddings is then fed into the transformer encoder which is made up of a Multi-Head Self-attention layer (MSP), a Multi-Layer Perceptron (MLP), and layer normalizations. The multiple attention heads allow the model to capture a richer set of features and relationships between image patches and attend to the multiple parts of the image simultaneously. The output of the final encoder layer is a sequence of contextualized embeddings that is then pooled and fed into the classification head, typically a single fully connected layer to obtain the class prediction.

Federated Learning is a machine learning technique that allows data to be trained on a distributed network of devices without the need for centralized data storage. This approach allows data to be kept private as no data is shared between devices and clients. The process typically starts with the server-side initialization of the model and subsequent distribution of the model parameters to the clients. Each device then trains the model using its locally stored data, updating the model parameters. These updates or parameters are then shared back with the central server for aggregation, combining the updated parameters from each device. Once the central model has been updated, it is once again reshared with the devices for further training with each training round further improving the model. Federated averaging (FedAvg) is an algorithm to aggregate the updates from the client devices. During each training iteration, clients compute the gradients using batches of the local data and update the local model weights. These weights are then shared with the central server where they are averaged [4, 5].

The described approach is categorized as a centralized federated learning strategy, where a central server coordinates the client devices and aggregates model updates during the training process. This can pose a challenge as there is a single point of failure in the central server and also limited scalability as the central server must be able to handle an increasingly large number of incoming updates as the number of devices increases. In a decentralized federated learning strategy, model updates are exchanged between devices without the orchestration from a central server. Devices may exchange local model updates with a subset of other devices to form a global model or until a consensus is reached. However, this is much more challenging to implement. In either case, FL offers major benefits with regard to privacy as sensitive data can remain on the user's device but still be used for training. There may still be challenges such as with the heterogeneous nature of the data (non-IID), bad actors, and heterogeneous clients with different hardware.

3 Related Work

Numerous studies have examined the comparisons between vision transformers and CNN models. Table 1 [6] presents an overview of the properties and performance of various models on the ImageNet dataset. Generally, ViTs demonstrate similar or better top-1 accuracy compared to their CNN counterparts, although they tend to have lower throughput.

After the introduction of ViTs in 2019 and their subsequent dominance in the image

Model	Params (M)	FLOPs (B)	Throughput (image/s)	Top-1 (%)
CNN				
ResNet-50	25.6	4.1	1226	79.1
ResNet-101	44.7	7.9	753	79.9
ResNet-152	60.2	11.5	526	80.8
EfficientNet-B0	5.3	0.39	2694	77.1
EfficientNet-B1	7.8	0.70	1662	79.1
EfficientNet-B2	9.2	1.0	1255	80.1
EfficientNet-B3	12	1.8	732	81.6
EfficientNet-B4	19	4.2	349	82.9
Pure Transformer				
DeiT-Ti	5	1.3	2536	72.2
DeiT-S	22	4.6	940	79.8
DeiT-B	86	17.6	292	81.8
T2T-ViT-14	21.5	5.2	764	81.5
T2T-ViT-19	39.2	8.9	464	81.9
T2T-ViT-24	64.1	14.1	312	82.3
PVT-Small	24.5	3.8	820	79.8
PVT-Medium	44.2	6.7	526	81.2
PVT-Large	61.4	9.8	367	81.7
TNT-S	23.8	5.2	428	81.5
TNT-B	65.6	14.1	246	82.9
CPVT-S	23	4.6	930	80.5
CPVT-B	88	17.6	285	82.3
Swin-T	29	4.5	755	81.3
Swin-S	50	8.7	437	83.0
Swin-B	88	15.4	278	83.3

Table 1: CNN and Transformer model properties and performance on the ImageNet dataset. [6]

classification field, ConvNeXt was introduced in 2022 to revisit and modernize the idea of the convolutional network [7]. It took lessons from ViTs to create a new convolutional architecture that outperformed prior ViT-based models [7]. However, ConvNeXt's size proved to be inhibitory for training on edge devices with limited memory capacity. Thus, distilled versions of ConvNeXt were also introduced such as ConvNeXt-tiny which is comparable in size to the original ViTs (see Table 2).

ConvNeXts utilize similar training techniques as ViTs, including image augmentation, regularization, and the AdamW optimizer. Compared to ResNet, the sliding windows are modified to be non-overlapping similar to the image patches in Transformers, GELU is used instead of RELU, and layer normalization is used instead of batch normalization. The vast majority of comparisons and optimization between these and other models have been done in a standalone environment. In the context of federated learning environments, there has been significantly less work dedicated to investigating and comparing the performance of various models in terms of feasibility and optimization [8].

4 Motivation/Goal

The main goal of this project is to investigate the question: "How well would vision transformers perform in the context of federated learning?". With the impressive performance and various benefits provided by ViTs, looking at ViTs under a FL environment could help address problems around:

- Performance: ViTs are shown to perform comparably or better than other models with less computation [3]
- Scalability: Collecting data around a centralized location is challenging, especially data that are naturally decentralized or have a need to be such as medical imaging.
- Privacy: Multiple parties could collaborate and train the model without exposing any data.
- Exponential growth of model sizes: Other related models (CNNs, Convnext) are rapidly growing in size, becoming problematic on edge devices [9].
- Non-IID data: Different architecture could be more robust under Non-IID settings.

5 Characterization/Ideas

To test ViT performance under FL, a pipeline consisting of integrating the vision transformer models in a federated learning environment was built. There are various available tools for building the ViT models and the FL framework such as PyTorch and TensorFlow for model building, and FedCV, TensorFlow Federated (TFF), and Flower for FL. They each differ based on the level of customizability and simplicity of implementation. For example, the TFF has a built-in CIFAR-100 dataset already processed for FL and also integrates well with ML models from Keras or TensorFlow. The tools and frameworks chosen for this project were selected based on familiarity and ease of implementation. As this experiment required a bit more control and customizability, the ViT model was built independently with PyTorch, then integrated into an FL environment setup and simulated using Flower which has an emphasis on simplicity.

The CIFAR dataset is a collection of labeled images commonly used for ML image classification tasks and training. CIFAR-10 consists of 60,000 32x32 color images in 10 classes with 6,000 images per class. Classes include airplanes, automobiles, birds, cats, deer,

Model Name	Model Card	Num of Params
ViT	google/vit-base-patch16-224	87M
DeiT	facebook/deit-base-distilled-patch16-224	87M
DeiT-S	facebook/deit-small-distilled-patch16-224	22M
BiT-50	google/bit-50	25M
ConvNeXt	facebook/convnext-tiny-224	28M

Table 2: List of models tested as well as their number of parameters. The first three models are transformer-based models, while the last two primarily use convolutions. Note that DeiT-S was selected due to its comparable size to the distilled ConvNeXt.

dogs, frogs, horses, ships, and trucks. The dataset is split into 50,000 training images and 10,000 testing images. CIFAR-100 consists of 60,000 32x32 color images in 100 classes with 600 images per class. The classes are grouped into 20 superclasses, each containing 5 subclasses. For example, the "aquatic mammals" superclass contains the subclasses "beaver," "dolphin," "otter," "seal," and "whale." The dataset is split into 50,000 training images and 10,000 testing images. Both CIFAR-10 and CIFAR-100 are widely used in research as benchmark datasets for image classification algorithms and thus are also used here.

The approach of this experiment can roughly be divided into:

1. Data pre-processing: Datasets (CIFAR-10 and CIFAR-100) are pre-processed, including generating patch embeddings from images and random splitting to simulate various clients
2. Build ViT model: Implement the encoder, MLP head, and optionally load pre-trained ImageNet weights
3. Build FL architecture: Implement the centralized server, FedAVG algorithm [4]. An overview of the FL architecture is provided in Figure 2 [10].
4. Train models in simulated FL environment
5. Performance Analysis: Assess feasibility and performance metrics

6 Evaluation Methodology

To evaluate the performance of transformer models under federated learning 5 models were selected to compare against: ViT (the original ViT base model), BiT-50 (the ResNet-50 model that was compared against in the original paper),

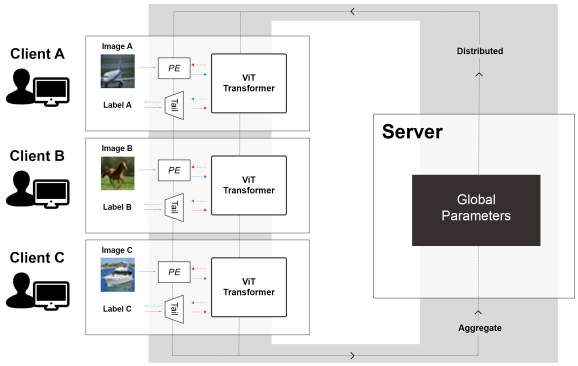

Figure 2: Overview of the implemented federated learning environment integrating vision transformers [10]

ConvNeXt-Tiny (a distilled version of the SOTA ConvNeXt model), DeiT (a data-efficient version of ViT-B), and DeiT-S (a distilled version of ViT that is comparable in size to ConvNeXt). The sizes and addresses to the model cards for these models can be found in Table 2.

Traditional datasets CIFAR-10 and CIFAR-100 were used. To test under a non-IID environment, the TensorFlow-Federated CIFAR-100 dataset was adapted to work with the PyTorch models [11]. Note that for the IID case, the data was normally distributed among clients. In terms of hardware, due to memory constraints everything was run on CPUs. FL environments were simulated using the FL framework, Flower, with an Intel i9-12900K and 32GB of DDR4 RAM. To mimic real-world settings with heterogeneous clients, a Windows Surface 2 laptop running on an Intel Core i7-8650U with 16GB RAM and an old Macbook Air running on an Intel Core i5-4250U with 4GB RAM was used with the Intel i9-12900k PC acting as the server.

In total, 3 different environments were explored: IID Simulated FL, Non-IID Simulated

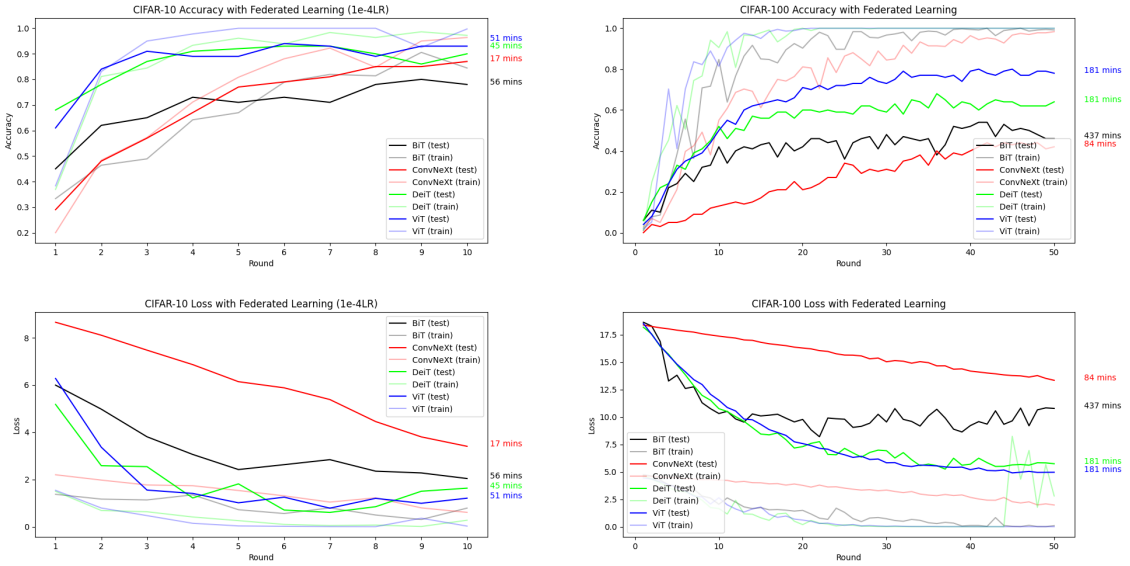

Figure 3: CIFAR-10 and CIFAR-100 accuracies and loss under the IID simulated FL setting. The Number of clients is 5 with 2 clients selected at each round for training. CIFAR-100 is trained for longer (50 rounds) in order to converge. All other parameters are identical. Solid lines are test metrics and faded lines are training metrics and the time elapsed for each model is shown on the right of each figure.

FL, and Non-IID Real FL. From these, training and testing metrics for accuracy and loss as well as the total runtime were gathered. Note that for the training metrics, the scores from all the clients were aggregated at the end of each round to get an average score which is the training metric for that round. Additionally, a held-out dataset was used to compute the testing metrics, server-side.

7 Results

7.1 IID Simulated FL

The CIFAR-10 and CIFAR-100 accuracies and loss under the simulated FL environment with IID data is shown in Figure 3. The evaluation of different models in this task revealed interesting insights into their performance. The ViT models (green and blue) in particular, consistently outperformed other models in terms of accuracy, indicating the effectiveness of their attention-based mechanism. However, the distilled ConvNeXt model was also found to have competitive performance, while having outstanding efficiency with runtimes of less than half the ViT models. The distilled ConvNeXt model may be a promising alternative that provides a good trade-off between performance and efficiency, making it an attractive option for real-world applications where computational resources are limited such as in an FL environment. In contrast, the BiT-50 model was found to perform the worst in terms of both runtimes and accuracy, as expected, highlighting the significant advancements ViT models have made.

7.2 Non-IID Simulated FL

Moving towards data that is more representative of a real-world setting, results from the non-IID data (Figure 4) reveal drastically worse overall performance compared to the IID case with notable overfitting issues, and a failure to achieve high accuracies across all models. It is clear that the non-IID data setting presents a more challenging scenario for the models. Nonetheless, both ViT models outperformed the ConvNeXt models with better accuracy and loss. This suggests that the attention mechanisms employed in the ViT architecture could be more effective in handling the challenges of non-IID data and can possibly capture the relevant features better.

7.3 Non-IID Real FL

Finally, under the most realistic setting with real non-simulated clients, Figure 5 again shows how ViTs excel under an FL setting. Note that due to time constraints, not all models

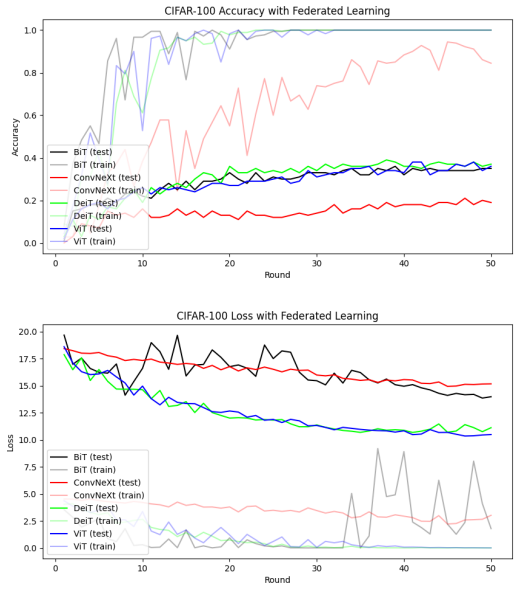

Figure 4: Non-IID Simulated model accuracy and loss using CIFAR-100 dataset from Tensorflow-Federated [11]

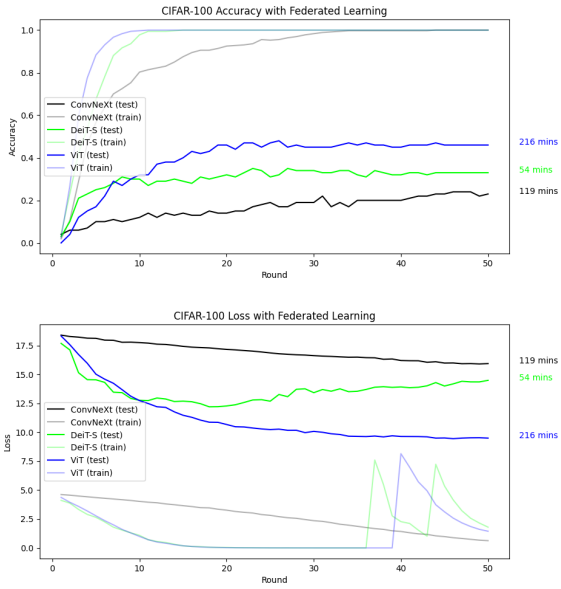

Figure 5: Accuracy and loss for models in Non-IID Real FL setting. Due to time constraints, only ConvNeXt, ViT, and DeiT-S are evaluated.

were tested under this setting. Additionally, DeiT was replaced by its smaller DeiT-S showing how even distilled versions of ViTs (less than half the size) outperform ConvNeXt. DeiT-S also has the best runtime, outperforming ConvNeXt under the simulated environment in Figure 3. Further research could be done to determine how well these models will perform on other non-IID datasets, as different non-IID patterns may require different learning strategies. Overall, these results highlight the importance of considering the distributional properties of the data when designing machine learning models and the potential benefits of using attention-based architectures like ViTs in both IID and non-IID scenarios under an FL environment.

8 Conclusion

In conclusion, this project aimed to investigate the performance of vision transformers (ViTs) in the context of federated learning (FL). The potential benefits of ViTs in FL were explored in terms of accuracy and computational efficiency. Using PyTorch and integrating the models into the FL environments with Flower, model performance was evaluated using CIFAR-10 and CIFAR-100 datasets.

Through the evaluation of multiple environments including IID Simulated FL, Non-IID Simulated FL, and Non-IID Real FL, it was found that ViT models consistently outperformed other models in terms of accuracy. In particular, the attention mechanisms employed in the ViT architecture were found to be more effective in handling the challenges of non-IID data. Although the ConvNeXt model showcased competitive performance and outstanding efficiency, distilled ViTs such as DeiT-S, which possess a similar number of parameters, offer an attractive alternative for real-world applications where computational resources may be limited. However, it is important to note that as models are made smaller, a trade-off occurs between accuracy and efficiency ultimately leading to deteriorating performance (see Figure 6 in the appendix).

Overall, the results of this study support the feasibility and performance of ViTs in the context of FL, while also highlighting the potential of distilled models in providing a balance between performance and efficiency. Further research could investigate the performance of other ViT variants like Swin-transformers under federated learning settings, and explore other FL techniques such as decentralized FL.

References

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural

network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[2] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[4] 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.

[5] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

[6] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[8] Xiaojiang Zuo, Qinglong Zhang, and Rui Han. An empirical analysis of vision transformer and cnn in resource-constrained federated learning. In *Proceedings of the 2022 5th International Conference on Machine Learning and Machine Intelligence*, pages 8–13, 2022.

[9] Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2574–2594, 2020.

[10] Sangjoon Park and Jong Chul Ye. Multi-task distributed learning using vision transformer with random patch permutation. *IEEE Transactions on Medical Imaging*, 2022.

[11] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

A Additional Figures

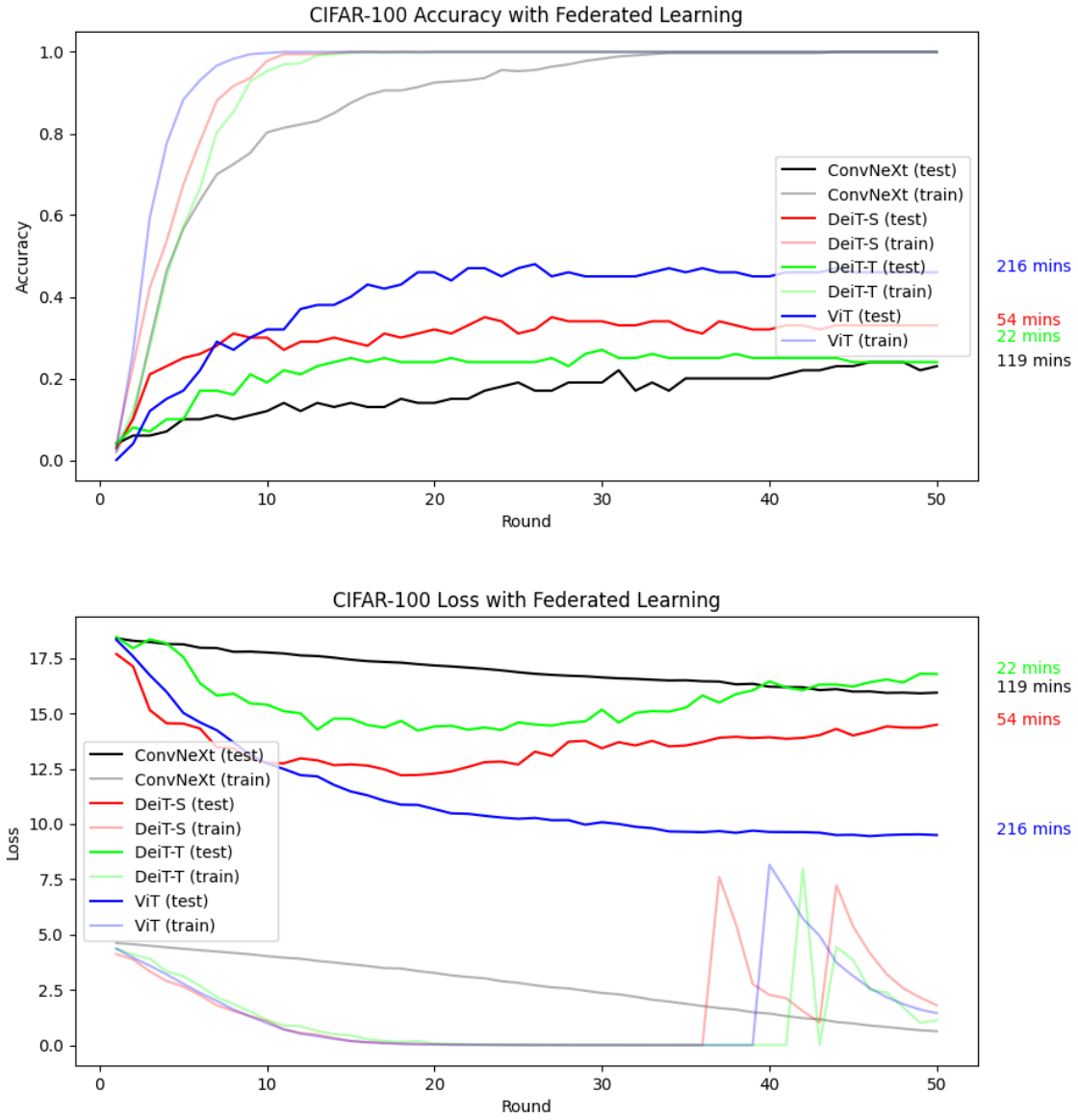

Figure 6: Additional results for the Non-IID Real FL setting. DeiT-T is the smallest distilled ViT model with only 5M parameters, yet is still competitive with ConvNeXt and runs substantially faster.