Seizing Critical Learning Periods in Federated Learning

Gang Yan[1], Hao Wang[2] and Jian Li[1]

[1] SUNY-Binghamton University
[2] Louisiana State University
gyan2@binghamton.edu, haowang@lsu.edu, lij@binghamton.edu

Abstract

Federated learning (FL) is a popular technique to train machine learning (ML) models with decentralized data. Extensive works have studied the performance of the global model; however, it is still unclear how the training process affects the final test accuracy. Exacerbating this problem is the fact that FL executions differ significantly from traditional ML with heterogeneous data characteristics across clients, involving more hyperparameters. In this work, we show that the final test accuracy of FL is dramatically affected by the early phase of the training process, i.e., *FL exhibits critical learning periods*, in which small gradient errors irrecoverably impact the final test accuracy. To further explain this phenomenon, we generalize the trace of Fisher Information Matrix (FIM) to FL and define a new notion called FedFIM, a quantity reflecting the local curvature of each client from the beginning of training in FL. Our findings suggest that the *initial learning phase* plays a critical role in understanding the FL performance. This is in contrast to many existing works which generally do not connect the final accuracy of FL to the early phase training. Finally, seizing critical learning periods in FL is of independent interest and could be useful for other problems such as the choices of hyperparameters including but not limited to the number of client selected per round, batch size, so as to improve the performance of FL training and testing.

Introduction

The ever-growing attention to data privacy and the popularity of mobile computing have impelled the rise of Federated learning (FL) (McMahan et al. 2017; Imteaj et al. 2021), a new distributed machine learning paradigm on decentralized data. A typical FL system consists a central server and multiple decentralized clients (e.g., smartphones and IoT devices). The central server initiates federated learning by sending a global model to clients. The clients then use their local data samples to train the received model with common deep learning algorithms and aggregate their local models to the central server. The central server updates the global model by aggregating the received local models and sends it to clients for further training. By repeating the local training and global aggregation, the central server obtains a global model jointly trained by decentralized clients without leaking any raw data. This unique distributed nature enables an extensive deployment of FL that trains deep learning models on sensitive private data, such as Google Keyboard (Yang et al. 2018).

The distributed nature of FL raises a series of new challenges in terms of *system performance* and *data statistics*. In FL systems, clients are typically loosely-connected mobile devices with limited communication bandwidth, computation power, and battery life. Unlike traditional centralized machine learning, data samples of each client in FL follow a non-identical and independent distribution (non-IID), introducing bias that slows down or even fails the training. A few recent studies have been proposed to address these challenges by model compression (Konečný et al. 2016; Suresh et al. 2017), communication frequency optimization (Wang and Joshi 2019; Wang et al. 2019; Karimireddy et al. 2020), and client selections (Lai et al. 2021; Wang et al. 2020a; Xiong, Yan, and Li 2021).

However, existing studies have not yet explored the significance of *critical learning periods*. Recent works have revealed that the first few training epochs—known as critical learning periods—determine the final quality of a deep neural network (DNN) model in traditional centralized ML (Achille, Rovere, and Soatto 2019; Jastrzebski et al. 2019; Golatkar, Achille, and Soatto 2019; Jastrzebski et al. 2021). During a critical period, deficits such as low quality or quantity of training data will cause irreversible model degradation, no matter how much additional training is performed after the period. The existence of critical periods in FL remains an open question due to the unique distributed nature of FL.

In this paper, we seek critical learning periods in FL with systematic experiments and theoretical analysis, and we emphasize the necessity of seizing the critical learning periods to improve FL training efficiency. Specifically, through a range of carefully designed experiments on different ML models and datasets, we observe the consistent existence of critical learning periods in the FL training process. We further propose a new metric named Federated Fisher Information Matrix (FedFIM) to describe and explain this phenomenon. FedFIM is calculated based on a classical statistics notion of Fisher Information Matrix (FIM) (Amari and Nagaoka 2000) that approximates the local curvature of the loss surface in FL efficiently. We show that the phenomenon of critical learning periods in FL can be explained using the
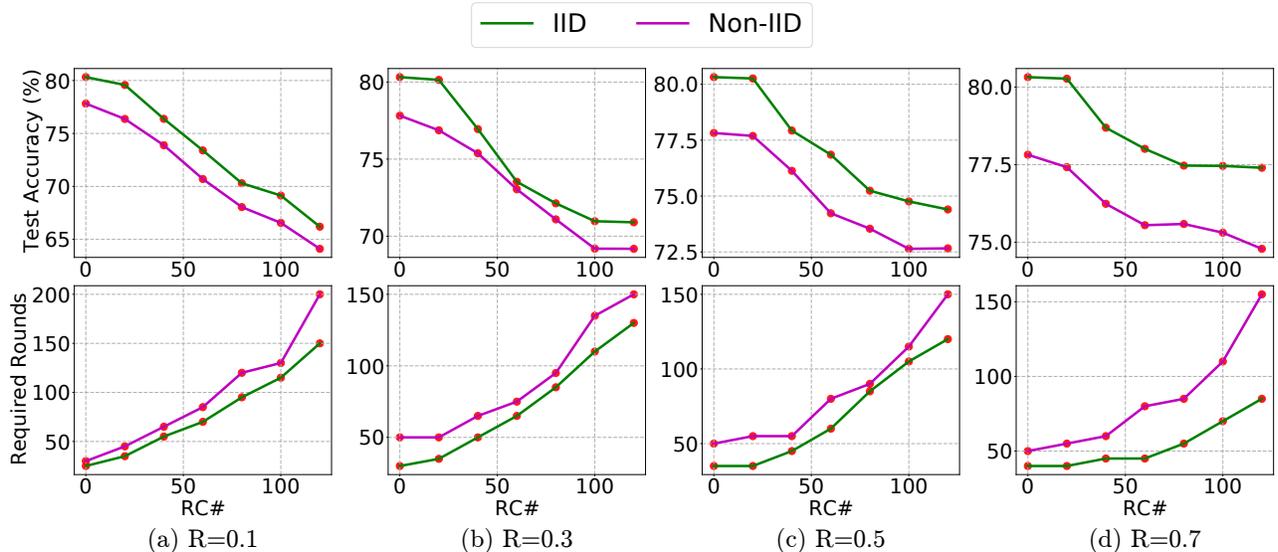
Figure 1: FL exhibits critical learning periods. (top) The final accuracy achieved by ResNet-18 on both IID and Non-IID CIFAR-10 with FedAvg using partial local datasets (where R indicates the ratio of local datasets) for training as a function of the communication round at which the partial training dataset is recovered to the entire training dataset. The test accuracy of FL is permanently impaired if the training dataset is not recovered to the entire training dataset early enough, no matter how many additional training rounds are performed. (bottom) Communication rounds vs. recover round (RC#). The total communication rounds required to achieve the corresponding final accuracy are significantly increased as a function of the recover round.

trace of FedFIM, a quantity reflecting the local curvature of each clients from the beginning of the training in FL. Our findings suggest that the *initial learning phase* plays a critical role in understanding the FL performance, complementing many existing studies that generally ignore the connection between the final model accuracy and the early phase training. To the best of our knowledge, this is the first work towards seizing critical learning periods in FL framework for training efficiency. Our main contributions are as follows:

1. We discover that critical learning periods consistently exist in FL with representative models and datasets through our carefully-designed experiments.

2. We systematically explore the impacts of critical learning periods for FL under a wide range of FL hyperparameters, including client availability, learning rates, batch size, and weight decay.

3. We propose a new notion dubbed Federated Fisher Information Matrix (FedFIM) and analyze the phenomenon of critical learning periods in FL through the trace of FedFIM. We show that model quality during critical periods correlates strongly with the trace of FedFIM.

Our experiment details, parameter settings and additional experimental results can be found in (Yan, Wang, and Li 2021).

Background

Federated Learning

The goal of FL is to solve a joint optimization problem as

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}, \mathcal{D}) := \sum_{j \in \mathcal{N}} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} \mathcal{L}_j(\mathbf{w}, \mathcal{D}_j), \quad (1)$$

where $\mathbf{w}$ denotes the model parameters, $\mathcal{N}$ denotes the set of clients, $\mathcal{D}_j$ is the local dataset of client $j \in \mathcal{N}$, the entire training dataset is $\mathcal{D} = \cup_{j \in \mathcal{N}} \mathcal{D}_j$, and $\mathcal{L}_j(\mathbf{w}, \mathcal{D}_j)$ is the local loss function of client j. A typical solution to this optimization problem is *federated averaging* (FedAvg) algorithm (McMahan et al. 2017). Specifically, FedAvg initializes with a random global model $\mathbf{w}_0$ and iterates the following two steps within each communication round $t = 1, \cdots, T$:

- **Local training.** The central server sends the goal model $\mathbf{w}_{t-1}$ to a randomly selected subset of clients $\mathcal{N}_t \subset \mathcal{N}$. Each selected client $j \in \mathcal{N}_t$ performs local training using its own dataset $\mathcal{D}_j$:

$$\mathbf{w}_{t,j}(k) \leftarrow \mathbf{w}_{t,j}(k-1) - \eta \nabla \mathcal{L}_j(\mathbf{w}_{t,j}(k-1), \mathcal{D}_j), \ (2)$$

where η is the learning rate and $k = 1, \cdots, K$ is the index of local iterations.

- **Global aggregation.** The central server obtains a new global model $\mathbf{w}_t$ by weighted-averaging the local models collected from the selected clients in round t:

$$\mathbf{w}_t \leftarrow \sum_{j \in \mathcal{N}_t} \frac{|\mathcal{D}_j|}{|\cup_{j \in \mathcal{N}_t} \mathcal{D}_j|} \mathbf{w}_{t,j}(K). \quad (3)$$
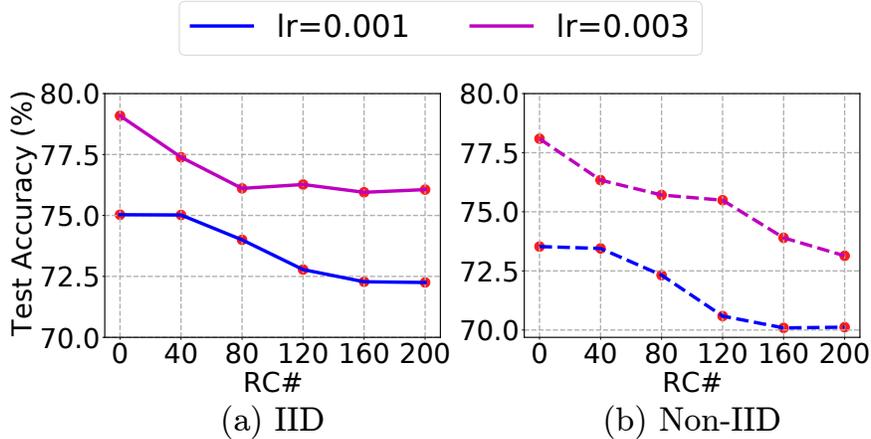
Figure 2: The existence of critical learning periods in FL: FedAvg trained on ResNet-18 using both IID and Non-IID CIFAR-10 with constant learning rates (lr).

There are a few variant federated learning algorithms, such as SCAFFOLD (Karimireddy et al. 2020), FedProx (Li et al. 2020a), and FetchSGD (Rothchild et al. 2020). We choose to perform observations and analysis based on FedAvg because its simplicity and generality extensively reduce the uncertainty of critical periods.

Critical Learning Periods

Critical learning periods were originally observed in early post-natal development of humans and animals that sensory deficits will cause lifelong irreversible skill impairment. Recently, researchers observed similar phenomenons in centralized deep learning that training a model with defective data such as blurred images in early epochs will decrease its final accuracy, no matter how many additional training epochs are performed (Agarwal et al. 2021; Achille, Rovere, and Soatto 2019; Jastrzebski et al. 2019; Golatkar, Achille, and Soatto 2019; Jastrzebski et al. 2021).

However, observing and justifying critical learning periods in FL are hindered a few obstacles: (i) FL involves multiple deep learning processes across randomly selected clients with their own data; (ii) the global model aggregated by local models at the central server has no direct information about the training data decentralized across clients; and (iii) FL has far more hyperparameters (e.g., the number of selected clients and data distribution) than centralized training that make it complicated to induce critical learning periods.

Critical Learning Periods in FL

We hypothesize that the final accuracy of FL is significantly affected by the initial learning phase, which we term as the critical learning periods in FL. Consider a model with loss function $\ell(\boldsymbol{x}; \boldsymbol{w})$, where ℓ reaches a minimum loss ℓ_{loss} with a test accuracy ℓ_{acc} when optimized with FedAvg across N decentralized clients on the entire training dataset $\mathcal{D}$. In addition, consider optimizing FedAvg across all clients only with a subset of the local training dataset $\mathcal{D}'_j \subset \mathcal{D}_j, \forall j \in \mathcal{N}$ in the first M communication rounds and then using the entire training dataset $\mathcal{D}$ afterwards. Then ℓ reaches a mini-

mum loss of $\ell'_{\mathrm{loss}}(M)$ with a test accuracy of $\ell'_{\mathrm{acc}}(M)$. The critical learning periods articulate that there exist M_1 and M_2 such that $\ell'_{\mathrm{acc}}(M_1) \geq \ell'_{\mathrm{acc}}(M_2)$ when $M_1 \leq M_2$, i.e., the initial learning phase is critical in determining the final performance of FL, and the effect of insufficient training (i.e., only using part of the entire training dataset) during the critical learning periods cannot be overcome, no matter how much additional training is performed.

In this section, we address two key questions pertains to the phenomenon of critical learning periods in FL. We first show via an extensive set of experiments that the critical learning periods can be observed across different popular ML models and datasets. We then reveal that the critical learning periods in FL stay robust under various training schemes.

FL Exhibits Critical Learning Periods

We perform extensive simulations using two representative ML models: ResNet-18 (He et al. 2016) and CNN, on popular datasets CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton 2009). To present the existence of critical learning periods in FL, we adopt the standard FedAvg (McMahan et al. 2017) which requires the entire training dataset throughout the training process, as well as its performance when only a subset of the training dataset on each client is involved in the first M communication rounds at which the training dataset is recovered to the entire training dataset. We call M as the "Recover Round" and denote R as the ratio of local datasets involved in training. We consider a system with $N = 64$ clients and FedAvg randomly selects a subset of 12 clients in each round. The batch size is of 16; the initial learning rate is set to 0.01 with a decay of 0.97 per round; and the SGD solver is adopted using an exponential annealing scheduling for the learning rate with a weight decay of 5×10^{-4}.

Figure 1 (top) reports the final performance of FL affected by the partial training datasets with different ratios R as a function of the recover round M. All results consistently endorse that the critical learning periods exist across all settings with different ratios of local datasets involved in the
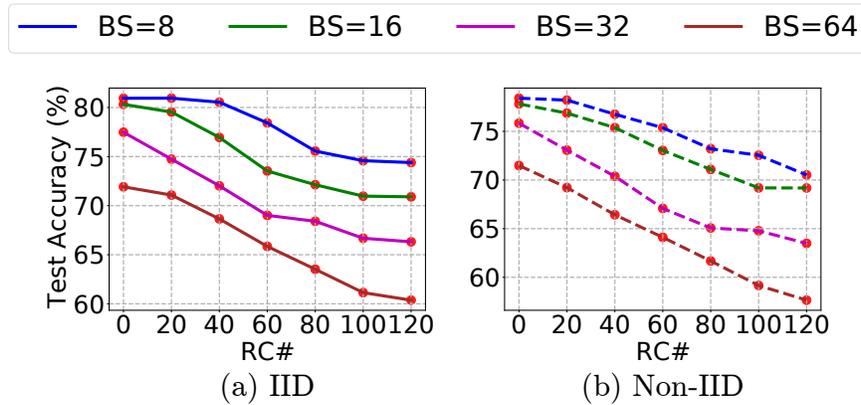
Figure 3: The existence of critical learning periods in FL: FedAvg trained on ResNet-18 using both IID and Non-IID CIFAR-10 with different batch sizes (BS).
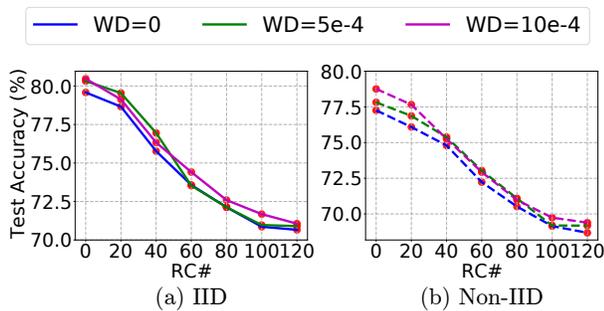

Figure 4: The existence of critical learning periods in FL: FedAvg trained on ResNet-18 using both IID and Non-IID CIFAR-10 with different weight decays (WD).

early learning phase: if the training dataset is not recovered to the entire dataset, at as early as the 20-th communication rounds, the final test accuracy of FL is severely degraded compared to the standard FedAvg. Comparing among different ratios R of local datasets involved in early training phase, it is not too surprising to see that lower R of local datasets in the early training phase makes drawing critical learning periods easier.

We further measure the total communication rounds required to achieve the corresponding final accuracy as a function of the recover round, as illustrated in Figure 1 (bottom). It is obvious that the communication rounds are significantly increased with a lower final test accuracy as a function of the recover round M. This further indicates the importance of the initial learning phase in determining the final performance of FL.

Learning Rate Annealing and Batch Size

We conduct the same experiments as in Figure 1 but using a constant learning rate rather than an annealing scheme. In particular, we set the constant learning rates to be 0.001 and 0.003, respectively. From Figure 2, we still observe the existence of critical learning periods in FL even with constant learning rates. Therefore the phenomenon of critical learn-

ing periods in FL are not resultant from an annealed learning rate in later rounds, and cannot be solely explained in terms of the loss landscape of the optimization in (1). Analogous results illustrating the impact of batch size are presents in Figure 3. Again the critical learning periods consistently exist regardless of the choice of batch size. This further suggests that the phenomenon of critical learning periods in FL cannot be simply explained by the differences in batch sizes.

Weight Decay

Similarly, the results for the same experiments as in Figure 1 but with different weight decays are presented in Figure 4. We still observe the critical learning periods as in Figure 1, but surprisingly the shapes of the critical learning periods are robust to the values of weight decays, i.e., changing the weight decays does not impact the shape of the critical learning periods.

Federated Fisher Information

Through extensive experiments, we have shown that the *initial learning phase* of the training process plays a critical role in the final test accuracy of FL. Our main contribution in this section is to show that this phenomenon can be explained by the trace of the *Federated Fisher Information Matrix (FedFIM)*, a quantity reflecting the local curvature of each clients from the beginning of the training in FL. We begin with the definition of the FIM for centralized training.

Fisher Information Matrix

Consider a probabilisitic classification model $p_{\mathbf{w}}(y|\boldsymbol{x})$, where $\mathbf{w}$ is the model parameter. Let $\ell(\boldsymbol{x}, y; \mathbf{w})$ be the cross-entropy loss function calculated for input $\boldsymbol{x}$ and label y. Denote the corresponding gradient of the loss computed for an example $(\boldsymbol{x}, y)$ as $g(\boldsymbol{x}, y; \mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \ell(\boldsymbol{x}, y; \mathbf{w})$. Then the Fisher Information Matrix (FIM) $\boldsymbol{F}$ for centralized training is defined as

$$\boldsymbol{F}(\mathbf{w}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}, \hat{y} \sim p_{\mathbf{w}}(y|\boldsymbol{x})}[g(\boldsymbol{x}, \hat{y})g(\boldsymbol{x}, \hat{y})^{\mathsf{T}}], \quad (4)$$

where the expectation is often approximated using the empirical distribution $\mathcal{X}$ induced by the centralized training
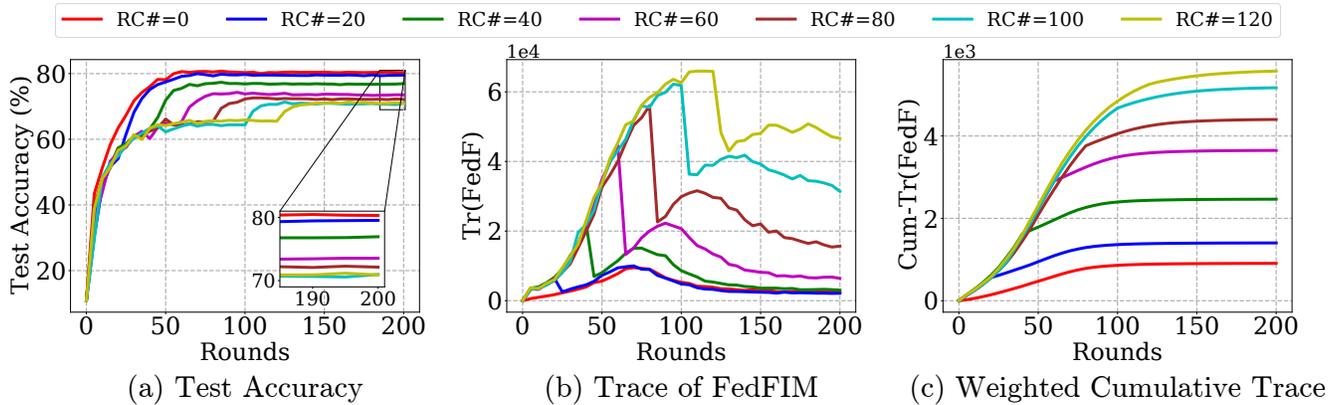
Figure 5: Connections between critical learning periods in FL and the Federated Fisher information achieved by ResNet-18 on IID CIFAR-10 with FedAvg using 30% of local datasets for training initially and recover to the entire datasets upon the recover round. (a) Test accuracy vs. recover rounds: the final test accuracy is permanently impaired if the training dataset is not fully recovered at as early as the 20-th round. (b) Trace of FedFIM vs. recover round. There exists a sharp increase of the trace of FedFIM in the early training phases. (c) Weighted cumulative sum of the trace of FedFIM vs. recover round.

dataset. Note that the FIM can be viewed as a local metric on how much the perturbation of the weights affects the network output (Amari and Nagaoka 2000). The FIM can also be seen as an approximation to the Hessian of the loss function (Martens 2014), and hence of the curvature of the loss landscape at a particular point $\mathbf{w}$ during training. This provides a natural connection between the FIM and the optimization procedure (Amari and Nagaoka 2000).

However, the computation of FIM in (4) requires the availability of the entire training dataset for the global model at the server. Unfortunately, this is infeasible for FL since training data is decentralized across clients. Hence we cannot compute FIM for FL as in (4). We now introduce a new notion to overcome this challenge.

Federated Fisher Information Matrix

Given that training data resides in each client, and the training process of FL in (2) and (3), we first introduce the notation of $\mathbf{F}_j(\mathbf{w})$, which represents the local FIM on client $j \in \mathcal{N}$:

$$\mathbf{F}_j(\mathbf{w}) = \mathbb{E}_{\boldsymbol{x}_j \sim \mathcal{X}_j, \hat{y}_j \sim p_{\mathbf{w}}(y_j|\boldsymbol{x}_j)}[g(\boldsymbol{x}_j, \hat{y}_j)g(\boldsymbol{x}_j, \hat{y}_j)^{\intercal}], \quad (5)$$

where $\mathcal{X}_j$ is the empirical distribution induced by the local dataset $\mathcal{D}_j$ of client j. Note that $\mathbf{F}_j(\mathbf{w})$ is computed using the global model $\mathbf{w}$ on the local dataset $\mathcal{D}_j$, and can be considered as a local metric measuring how the perturbation of the global model affects the FL training performance from the perspective of client j. As a result, the overall impact of the perturbation of the global model on the final output, which we define as the Federated Fisher Information Matrix (FedFIM) $\mathbf{FedF}$ for FL, can be computed using the weighted average of local FIM across all clients:

$$\mathbf{FedF}(\mathbf{w}) = \sum_j \frac{|\mathcal{D}_j|}{|\mathcal{D}|} \mathbf{F}_j(\mathbf{w}), \quad (6)$$

where the weight of client j is the size of its dataset. The rationale is that lower local FIM often has little effect on

the final performance. We denote the trace of $\mathbf{FedF}$ as $\mathrm{Tr}(\mathbf{FedF})$.

Experimental Results

We conduct similar experiments as in Figure 1 with partial local datasets involved in the initial learning phases and the training datasets recover to the entire datasets at the "recover rounds" (RC#). The test accuracy and the trace of FedFIM with different recover rounds and $R = 0.3$ on IID and Non-IID CIFAR-10 are presented in Figures 5 and 6.

First of all, we again observe the existence of critical learning periods since if the training dataset is not recovered to the entire datasets, e.g., at as early as the 20-th communication rounds, the final test accuracy of FL is permanently impaired. Second, this information is fully reflected via the trace of FedFIM as shown in Figure 5 (b) for IID case and Figure 6 (b) for Non-IID case. We observe a sharp increase in the trace of the FedFIM in the early phases of the FL training process, which coincides with dramatic increase of the test accuracy in the early training phase. The information starts to decrease when the test accuracy starts to plateau. Since the training datasets are recovered from 30% of local datasets to the entire datasets at the recover rounds, additional data further boosts the test accuracy as shown in Figure 5 (a). However, such a test accuracy boost decreases significantly as the recover rounds increase. This further suggests that the initial learning phases play a critical role in the FL performance and the permanent model degradation is irreversible no matter how much additional training is performed after the critical learning periods. Correspondingly, the accuracy boosting results in a slight increase in the trace of FedFIM, and the information decreases again when the test accuracy starts to plateau.

In general, the measures of test accuracy and trace of FedFIM are noisy, especially with Non-IID dataset as shown in Figure 6. This is because for instance the learning rate has to be adjusted in order to compensate for possible gen-

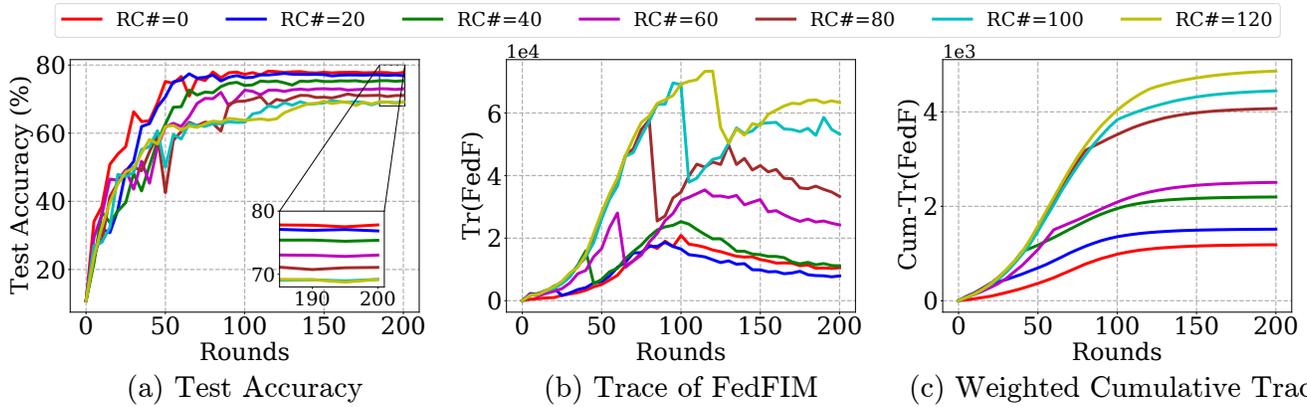(a) Test Accuracy   (b) Trace of FedFIM   (c) Weighted Cumulative Trace

Figure 6: Connections between critical learning periods in FL and the Federated Fisher information achieved by ResNet-18 on Non-IID CIFAR-10 with FedAvg using 30% of local datasets for training initially and recover to the entire datasets upon the recover round.

eralization issues of the training process (Jastrzebski et al. 2017; Smith et al. 2018). To this end, we further consider a weighted cumulative sum of the trace of FedFIM as follows

$$\text{Cum-Tr}(\boldsymbol{FedF})(k) = \sum_{i=0}^{k} \eta_i \text{Tr}(\boldsymbol{FedF}_i), \qquad (7)$$

where η_i is the learning rate at the i-th round, and $\boldsymbol{FedF}_i$ is the Federated Fisher Information Matrix at the i-th round. The trace of FedFIM represents the degree of whether the local data is good enough to improve the model. A larger values correspond to less model information. This is exactly observed in Figure 5 (c) and Figure 6 (c), where a late recovery results in larger weight cumulative trace.

Seizing Critical Learning Periods

We use carefully-crafted experiments to evaluate the idea that seizes critical learning periods to improve the FL training efficiency, though existing literature largely ignore the critical learning periods in FL training process. The experiments run on PyTorch on Python 3 with NVIDIA RTX 3060 GPU. The total number of clients is 25 and a subset of 25 clients are randomly selected in each round.

Specifically, we train ResNet-18 on IID and Non-IID CIFAR-10 with FedAvg under different settings as shown in Figures 7 and 8:

- **All Clients**: All clients participate in federated learning.
- **Partial Clients**: Only a subset of the clients (e.g., 60%) participate in federated learning.
- **All Clients in critical periods else Partial Clients**: All clients participate in training during the critical learning periods. After that, only a subset of clients (e.g., 60%) remain in training.
- **All Data**: Each client processes all data in local training.
- **Partial Data**: Each client processes only partial local datasets (e.g., 25%) in local training.
- **All Clients in critical periods else Partial Clients**: Each client uses its entire local dataset for training during the

critical learning periods, and only uses their partial local dataset afterwards.

By seizing the critical periods in FL, we summarize the counter-intuitive experimental results as follows:

No need to involve all clients in training all along. The conventional FedAvg requires the entire training datasets across all clients throughout the training process. However, some clients may not be available for training, e.g., due to unreliable network connection. To illustrate the impact of critical learning periods, we further consider a heuristic in which all clients are involved in the training during the critical learning periods and then only a subset of clients (e.g., 60%) are involved afterwards.

Figure 7(a) and Figure 8(a) show the test accuracy v.s. wall-clock time. There exists a requirement on the number of clients involved in training which provides similar test accuracy as using all clients (FedAvg) throughput. For example, with all clients participate in the FL training during the critical learning periods, and then only 60% of clients afterwards, the final test accuracy is similar to that using all clients throughout the training process. Hence there is no need to involve all clients throughout the training process. Figure 7(b) and Figure 8(b) show the train loss v.s. wall-clock time. The participated client number requirement reduces the training time than using all clients (FedAvg) throughput. It is clear that leveraging critical learning periods for FL training, even in a heuristic manner, can significantly improve the training efficiency with a reduced training time while maintaining final test accuracy.

No need to train a model with all local data for each client. We consider the challenge that FL clients have heterogeneous system capabilities, e.g., can only process part of the local data for training. We use a heuristic with entire local datasets used for training during the critical learning periods and then only partial local datasets involved afterwards.

Figure 7(c) and Figure 8(c) show the test Accuracy v.s. wall-clock time. There exists a training dataset requirement which provides similar test accuracy as using the entire dataset (FedAvg) throughput. For example, with the entire
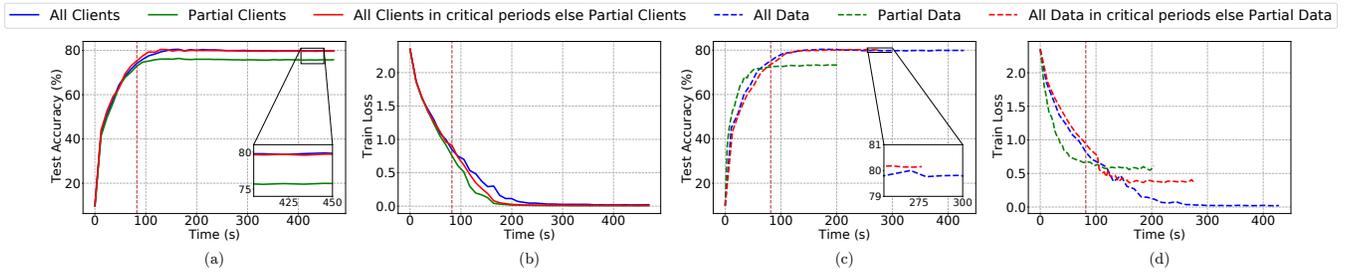
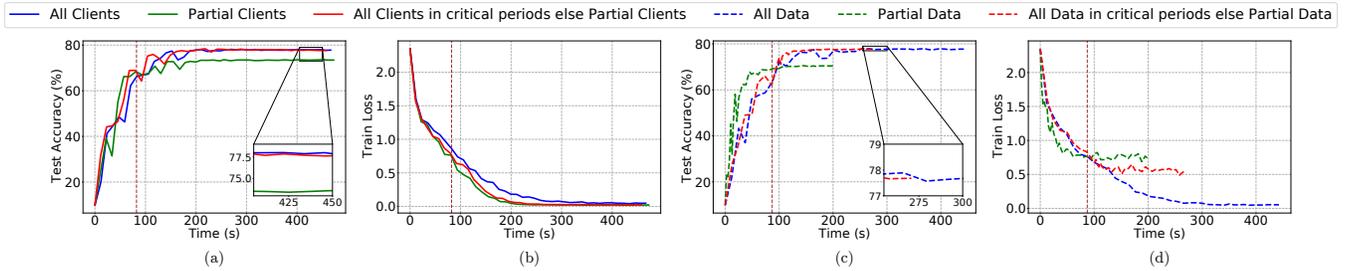Figure 7: Seizing the critical learning periods in FL training with ResNet-18 on IID CIFAR-10.

Figure 8: Seizing the critical learning periods in FL training with ResNet-18 on Non-IID CIFAR-10.

training datasets used in the FL training during the critical learning periods, and then only 25% of local datasets afterwards, the final test accuracy is similar to that using the entire datasets throughout the training process. Hence there is no need to use the entire training datasets throughout the training process. Figure 7(d) and Figure 8(d) present the train loss v.s. wall-clock time, the training dataset requirement (the heuristic) reduces the training time than using the entire dataset (FedAvg) throughput. Again, we observe that the early learning phase plays a critical role in FL performance and leveraging it can significantly improve the training efficiency of FL.

Overall, we can save 40%-50% of the training time and 50%-65% of the total clients but achieve a close final model accuracy when training ResNet-18 on the IID and non-IID CIFAR-10 dataset.

Related Work

Since the term of federated learning was introduced in the seminal work (McMahan et al. 2017), there is an explosive growth in federated learning research. For example, a line of works focuses on designing algorithms to achieve higher learning accuracy and analyze their convergence performance, e.g., (Smith et al. 2017; Li et al. 2020b; Liu et al. 2020; Wang et al. 2020b; Xiong, Yan, and Li 2021). Another line of works aim to improve the communication efficiency between the central server and clients through compressions or sparsification, (Konečnỳ et al. 2016; Suresh et al. 2017; Xu et al. 2019), communication frequency optimization (Wang and Joshi 2019; Wang et al. 2019; Karimireddy et al. 2020), client selections (Lai et al. 2021; Wang et al. 2020a), etc. Additionally, a lot of efforts have been put on exploring the privacy and fairness of federated learning (Bonawitz et al. 2017; Geyer, Klein, and Nabi 2017; Hitaj,

Ateniese, and Perez-Cruz 2017; Melis et al. 2019; Zhu, Liu, and Han 2019; Mohri, Sivek, and Suresh 2019; Wang et al. 2020b). These studies are often under the implicit assumption that all learning phases during the training process is equally importantly. Our work focuses on showing that the initial learning phase plays a critical role in the federated learning performance, which is orthogonal to the aforementioned studies.

Conclusion

The recent record-breaking development of machine learning (ML) algorithms, particularly in the area of deep neural networks (DNNs) motivates a tremendously growing demand of bringing DNN aided intelligence into modern ML applications. Different from conventional ML that needs to collect all training data in a centralized location, federated learning (FL) a promising paradigm that can obviate the need for centralized data. However, federated learning brings new challenges, for example, clients in FL are usually much more resource-constrained in terms of communication bandwidth, storage, computation power and more. Extensive works have focused on improving the efficiency of FL using compression, communication frequency optimization and so on. However, existing studies have not yet explored the significance of *critical learning periods* in FL.

In this paper, we seized the existence of critical learning periods in federated learning so as to improve the FL training efficiency. Though a range of carefully designed experiments on different ML models and datasets, we showed that critical learning periods consistently exists in the training process of FL. To explain such a phenomenon, we further proposed a new metric called Federated Fisher Information Matrix. Our findings suggest that the initial learning phase plays a critical role in the final performance of FL.

Acknowledgements

References

Achille, A.; Rovere, M.; and Soatto, S. 2019. Critical Learning Periods in Deep Networks. In *Proc. of ICLR*.

Agarwal, S.; Wang, H.; Lee, K.; Venkataraman, S.; and Papailiopoulos, D. 2021. ACCORDION: Adaptive Gradient Communication via Critical Learning Regime Identification. In *Proc. of MLSys*.

Amari, S.-i.; and Nagaoka, H. 2000. *Methods of Information Geometry*, volume 191. American Mathematical Soc.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *Proc. of ACM CCS*.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially Private Federated Learning: A Client Level Perspective. *arXiv preprint arXiv:1712.07557*.

Golatkar, A. S.; Achille, A.; and Soatto, S. 2019. Time Matters in Regularizing Deep Networks: Weight Decay and Data Augmentation Affect Early Learning Dynamics, Matter Little Near Convergence. *Proc. of NeurIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. of IEEE CVPR*.

Hitaj, B.; Ateniese, G.; and Perez-Cruz, F. 2017. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning. In *Proc. of ACM CCS*.

Imteaj, A.; Thakker, U.; Wang, S.; Li, J.; and Amini, M. H. 2021. A Survey on Federated Learning for Resource-Constrained IoT Devices. *IEEE Internet of Things Journal*.

Jastrzebski, S.; Arpit, D.; Astrand, O.; Kerg, G. B.; Wang, H.; Xiong, C.; Socher, R.; Cho, K.; and Geras, K. J. 2021. Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization. In *Proc. of ICML*.

Jastrzebski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three Factors Influencing Minima in SGD. *arXiv preprint arXiv:1711.04623*.

Jastrzebski, S.; Kenton, Z.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. J. 2019. On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length. In *Proc. of ICLR*.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proc. of ICML*.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint arXiv:1610.05492*.

Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. *Technical Report, University of Toronto*.

Lai, F.; Zhu, X.; Madhyastha, H. V.; and Chowdhury, M. 2021. Oort: Efficient Federated Learning via Guided Participant Selection. In *Proc. of USENIX OSDI*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated Optimization in Heterogeneous Networks. In *Proc. of MLSys*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020b. On the Convergence of FedAvg on Non-IID Data. In *Proc. of ICLR*.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2020. Federated Learning for Vision-and-Language Grounding Problems. In *Proc. of AAAI*.

Martens, J. 2014. New Insights and Perspectives on the Natural Gradient Method. *arXiv preprint arXiv:1412.1193*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. of AISTATS*.

Melis, L.; Song, C.; De Cristofaro, E.; and Shmatikov, V. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. In *Proc. of IEEE S&P*.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic Federated Learning. In *Proc. of ICML*.

Rothchild, D.; Panda, A.; Ullah, E.; Ivkin, N.; Stoica, I.; Braverman, V.; Gonzalez, J.; and Arora, R. 2020. FetchSGD: Communication-Efficient Federated Learning with Sketching. In *Proc. of ICML*.

Smith, S. L.; Kindermans, P.-J.; Ying, C.; and Le, Q. V. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In *Proc. of ICLR*.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. 2017. Federated Multi-Task Learning. In *Proc. of NIPS*.

Suresh, A. T.; Felix, X. Y.; Kumar, S.; and McMahan, H. B. 2017. Distributed Mean Estimation with Limited Communication. In *Proc. of ICML*.

Wang, H.; Kaplan, Z.; Niu, D.; and Li, B. 2020a. Optimizing Federated Learning on Non-IID Data With Reinforcement Learning. In *Proc. of IEEE INFOCOM*.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020b. Federated Learning with Matched Averaging. In *Proc. of ICLR*.

Wang, J.; and Joshi, G. 2019. Adaptive Communication Strategies to Achieve the Best Error-Runtime Trade-off in Local-update SGD. In *Proc. of SysML*.

Wang, S.; Tuor, T.; Salonidis, T.; Leung, K. K.; Makaya, C.; He, T.; and Chan, K. 2019. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE Journal on Selected Areas in Communications*, 37(6): 1205–1221.

Xiong, G.; Yan, G.; and Li, J. 2021. Straggler-Resilient Distributed Machine Learning with Dynamic Backup Workers. *arXiv preprint arXiv:2102.06280*.

Xu, Z.; Yang, Z.; Xiong, J.; Yang, J.; and Chen, X. 2019. ELFISH: Resource-Aware Federated Learning on Heterogeneous Edge Devices. *arXiv preprint arXiv:1912.01684*.

Yan, G.; Wang, H.; and Li, J. 2021. Critical Learning Periods in Federated Learning. *arXiv preprint arXiv:2109.05613*.

Yang, T.; Andrew, G.; Eichner, H.; Sun, H.; Li, W.; Kong, N.; Ramage, D.; and Beaufays, F. 2018. Applied Federated Learning: Improving Google Keyboard Query Suggestions. *arXiv preprint arXiv:1812.02903*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep Leakage from Gradients. *Proc. of NeurIPS*.