Resource-Constrained Federated Learning with Heterogeneous On-Device Models

Dapeng Oliver Wu

City University of Hong Kong

Outline

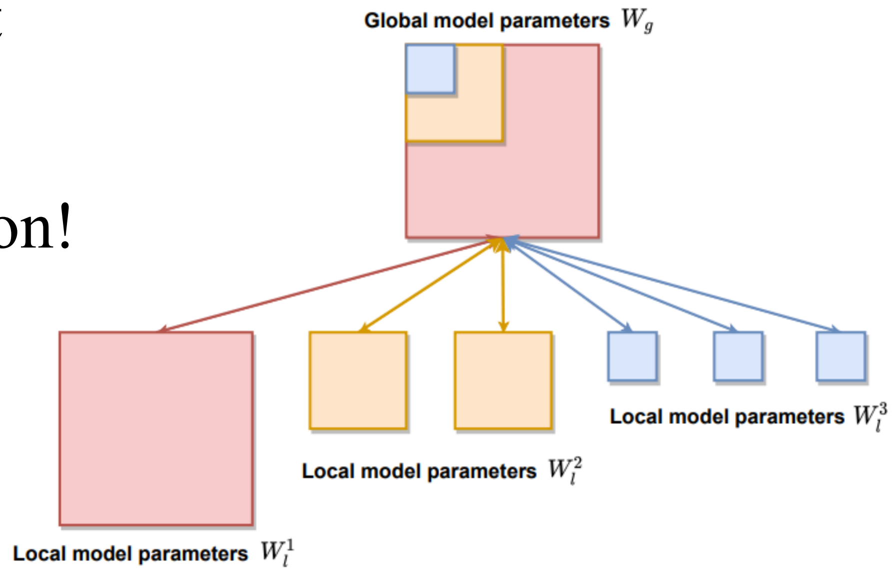
- Introduction
- Related Work
- FEDZKT: Federated Learning via Zero-shot Knowledge Transfer
- Experimental Evaluation
- Conclusion

Introduction

- Federated learning leverages on-device training at multiple distributed devices to obtain a knowledge-abundant global model without centralizing private on-device data.
- Classical federated learning algorithms, represented by FedAvg[2], require on-device training with the same model structure and size to perform the element-wise central average, which, however, impedes collaboration across heterogeneous hardware platforms.

Federated learning with heterogeneous on-device models

- HeteroFL
  - Assume the architecture of a small model can be a subnetwork of a large one, i.e., nesting structure
  - It is hard to find the architecture of MobileNet, i.e., a popular on-device model, as a subnetwork of other models, such as ShuffleNet and ResNet.
  - Model architecture limitation!



Diao, Enmao, Jie Ding, and Vahid Tarokh. "HeteroFL: Computation and communication efficient federated learning for heterogeneous clients." *ICLR 2021.*

Federated learning with heterogeneous on-device models

- FedMD, Cronus, FedH2L, FedDF
  - Design on-device models independently
  - Based on <u>federated distillation</u> technique
  - For personalization, security, and decentralization (logit information of on-device models), and for robust fusion (on-device model parameters), respectively

D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," arXiv preprint arXiv:1910.03581, 2019.
H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, "Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer," arXiv preprint arXiv:1912.11279, 2019.
Y. Li, W. Zhou, H. Wang, H. Mi, and T. M. Hospedales, "Fedh2l: Federated learning with model and statistical heterogeneity," arXiv preprint arXiv:2101.11296, 2021.
T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," arXiv preprint arXiv:2006.07242, 2020.

Federated learning with heterogeneous on-device models

- FedMD, Cronus, FedH2L, FedDF
  - Rely on certain prerequisites of on-device knowledge to extract and transfer knowledge
  - Assume a public dataset is available for knowledge transfer. But there may be a mismatch between the public dataset and the private data, resulting in poor knowledge transfer.

D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," arXiv preprint arXiv:1910.03581, 2019.
H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, "Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer," arXiv preprint arXiv:1912.11279, 2019.
Y. Li, W. Zhou, H. Wang, H. Mi, and T. M. Hospedales, "Fedh2l: Federated learning with model and statistical heterogeneity," arXiv preprint arXiv:2101.11296, 2021.
T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," arXiv preprint arXiv:2006.07242, 2020.

Our Design: FedZKT

- Independent on-device model design
- Data-free knowledge transfer
  - No need to have access to public data
    - We address it by zero-shot knowledge distillation
- Allow participation from resource-constrained and/or heterogeneous devices
  - We assign compute-intensive distillation task to a server

Related Work

1. Heterogeneous Federated Learning

- Data heterogeneity
- Device heterogeneity
    - Computing power or networking
    - E.g., address "straggler effect" introduced by some poorly performed devices; reduce local model size at all devices;
    - Most of these designs are still under the learning paradigm of FedAvg with homogeneous on-device models.

Related Work

2. Federated Distillation

- Model heterogeneity: FedMD, Cronus, FedH2L, FedDF (for personalization, security, decentralization, and robust fusion)
- Communication efficiency, privacy, data heterogeneity
- Weakness: assume a public dataset is available for knowledge transfer. But there may be a mismatch between the public dataset and the private data, resulting in poor knowledge transfer.

Related Work

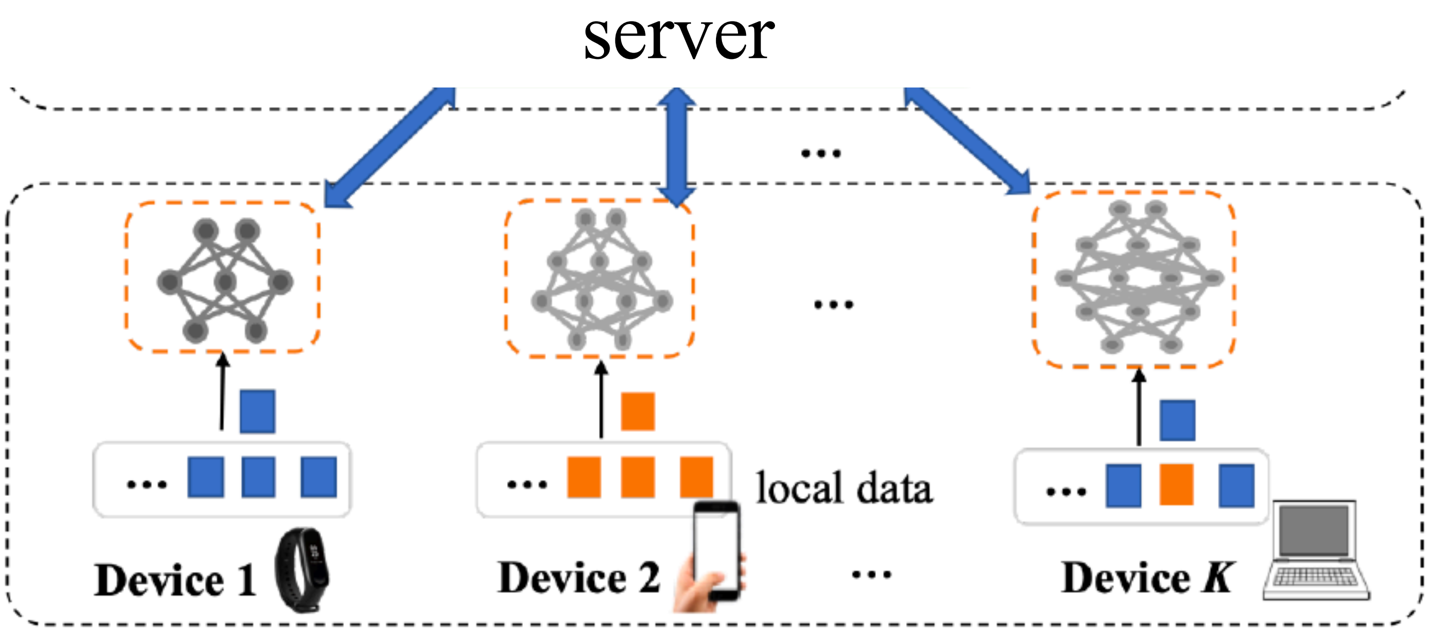
3. Data-Free Knowledge Distillation

- Typically, a generative model is learned to synthesize the queries that the student makes to the teacher.
  - E.g., model compression
  - Little attention to federated settings: **FeDGen**
    - Slow convergence due to data heterogeneity
    - Deploy generators on devices to augment local knowledge for data-free distillation

Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," arXiv preprint arXiv:2105.10056, 2021.

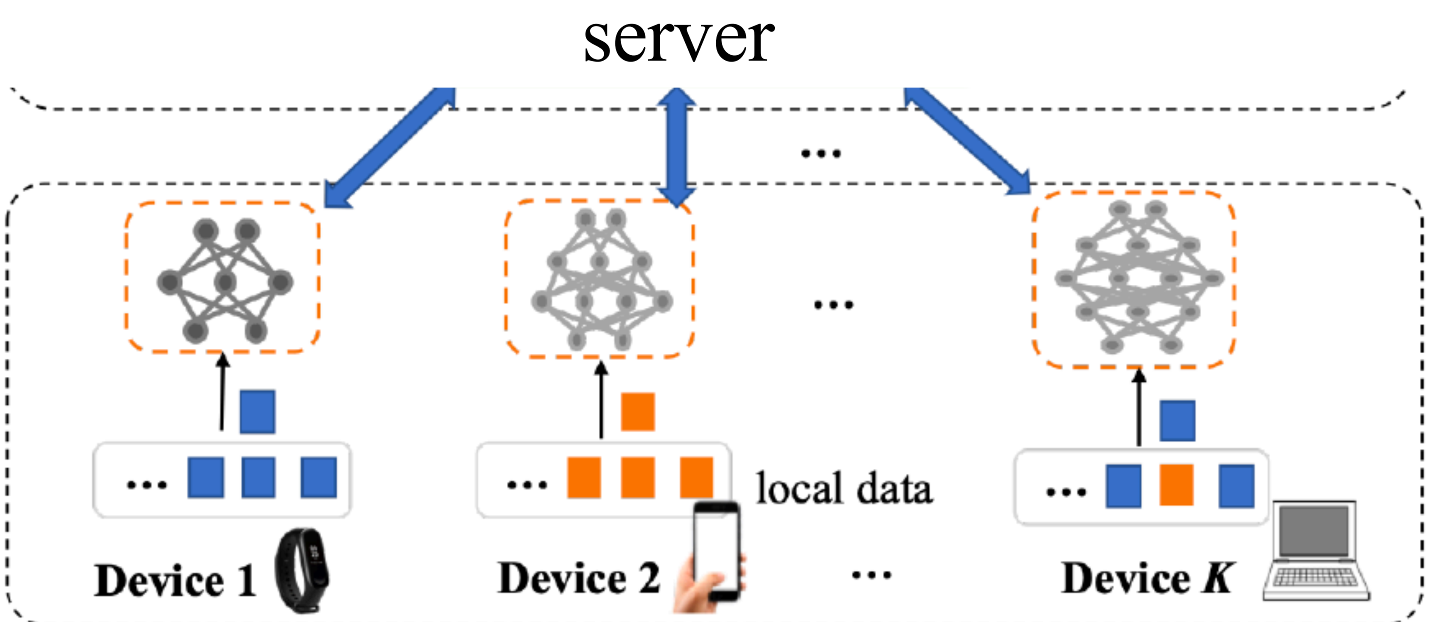
FedZKT: Federated Learning via Zero-shot Knowledge Transfer

- K heterogeneous devices (might be resource-constrained)
- A powerful server
- Goal: knowledge transfer in a data-free manner



FedZKT

- Unbalanced capabilities between server and devices
  - Assign the compute-intensive <u>zero-shot knowledge distillation</u> task to the server



Zero-Shot Knowledge Distillation

- Server's goal: obtain the global model to match the ensemble of on-device models without on-device data
- Intuitive idea: leverage a synthetic dataset to mimic local knowledge to minimize the loss of disagreement between the server (student) and device (teacher)

$$\min_{w} E_{x \sim \mathcal{D}_S} [\mathcal{L}(\mathcal{F}(x; w), f_{\text{ens}}(x))]$$

where $f_{\text{ens}}(x) = \frac{1}{|\mathcal{K}|} \sum_k f_k(x; w_k)$

Zero-Shot Knowledge Distillation

$$\min_{\mathcal{F}} \max_{G} E_{z\sim\mathcal{N}(0,1)}[\mathcal{L}(\mathcal{F}(G(z)), f_{\text{ens}}(G(z)))],$$

- Generative model G
  - Responsible to provide difficult inputs for the training of global model F
  - Maximizes the disagreement between the current global and on-device models
- Global model F
  - Matching knowledge at devices

Zero-Shot Knowledge Distillation

$$\min_{\mathcal{F}} \max_{G} E_{z \sim \mathcal{N}(0,1)}[\mathcal{L}(\mathcal{F}(G(z)), f_{\text{ens}}(G(z)))],$$

- Loss function L
  - Measure the disagreement between the global model and the on-device model ensemble
  - The key to distillation performance
    - The gradients computed through F and f_{ens} can easily impede the convergence of the optimizer, such as leading to gradient vanishing

Loss function

Kullback–Leibler (KL) divergence

$$\mathcal{L}_{\mathrm{KL}}(x) = \sum \mathcal{F}(x) \log \frac{\mathcal{F}(x)}{f_{\mathrm{ens}}(x)}.$$

- Tend to suffer from gradient vanishing[1] with respect to input data x when the student model (F) converges to the teacher model (f_{ens}).
- The problem becomes even more serious in zero-shot distillation settings, since the gradient vanishing will further affect the training of the generative model G.

[1]G. Fang, J. Song, C. Shen, X. Wang, D. Chen, and M. Song, "Data-free adversarial distillation," arXiv preprint arXiv:1912.11006, 2019

Loss function

l_1 norm loss

$$\mathcal{L}_{\ell_1}(x) = ||u(x) - \frac{1}{|\mathcal{K}|}\sum_k v_k(x)||_1,$$

- Compare the logit outputs (model outputs before the softmax layer) between the teacher and student models
- Lead to unstable training due to the large gradients
  - Federated learning requires aggregating distributed knowledge from participating devices.
  - Given diverse on-device model parameters, averaging logit values over on-device models may increase the gradients, making the whole learning process unstable.

Loss function

A new loss function: *softmax l_1* (SL) norm loss

$$\mathcal{L}_{\mathrm{SL}}(x) = ||\mathcal{F}(x) - f_{\mathrm{ens}}(x)||_1.$$

- Overcome the drawbacks of using KL-divergence loss and ℓ_1 norm loss
  - Two hypotheses
  - Empirical results

Loss function

Hypothesis 1. *When the global model F converges to the ensemble of on-device models f_{ens}, the gradients of KL divergence loss with respect to the input data x are smaller than those of the SL loss:*

$$||\nabla_x \mathcal{L}_{\text{KL}}(x)|| \underset{F \to f_{\text{ens}}}{\leq} ||\nabla_x \mathcal{L}_{\text{SL}}(x)||. \tag{6}$$

- Hypothesis 1 suggests that the SL loss can reduce the gradient vanishing effect than the KL-divergence loss for better convergence in zero-shot distillation.

Loss function

Hypothesis 2. *When the global model F converges to the ensemble of on-device models f_{ens}, the gradients of the ℓ_1 norm loss with respect to the input data x are greater than those of the SL loss:*

$$||\nabla_x \mathcal{L}_{\ell_1}(x)|| \underset{F \to f_{\text{ens}}}{\geq} ||\nabla_x \mathcal{L}_{\text{SL}}(x)||. \tag{7}$$

- Hypothesis 2 suggests that the SL loss can make the training more stable compared to the $\ell 1$ norm loss.

Loss function

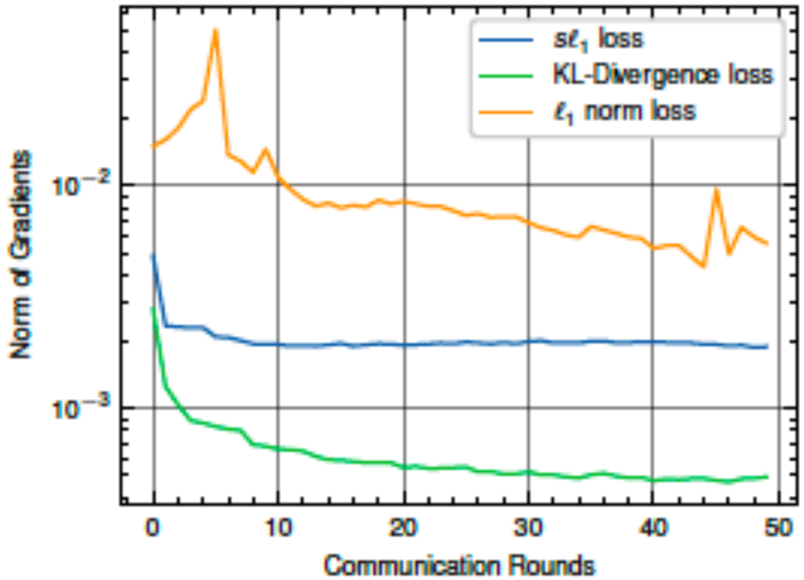


Fig. 2: **Norm of gradients w.r.t input data (MNIST, IID).** The gradients for the KL-divergence loss tend to vanish, while the gradients for the ℓ_1 norm loss are much larger and unstable during the learning process. The proposed SL loss overcomes both problems in the federated learning.

FedZKT

Bidirectional Knowledge Transfer

- Above design: knowledge transfer from devices to server
- Knowledge transfer from the server to devices
  - Intuitive idea: broadcast global model F

$$\min_{w_k} E_{x\sim\mathcal{D}_k}[\mathcal{L}(\mathcal{F}(x), f_k(x; w_k))]$$

  - Resource-constrained devices?
    - Run the round-trip distillation at the server
    - Reuse the well-learned generator G

$$\min_{f'_k} E_{z\sim\mathcal{N}(0,1)}[\mathcal{L}(\mathcal{F}(G(z)), f'_k(G(z)))].$$

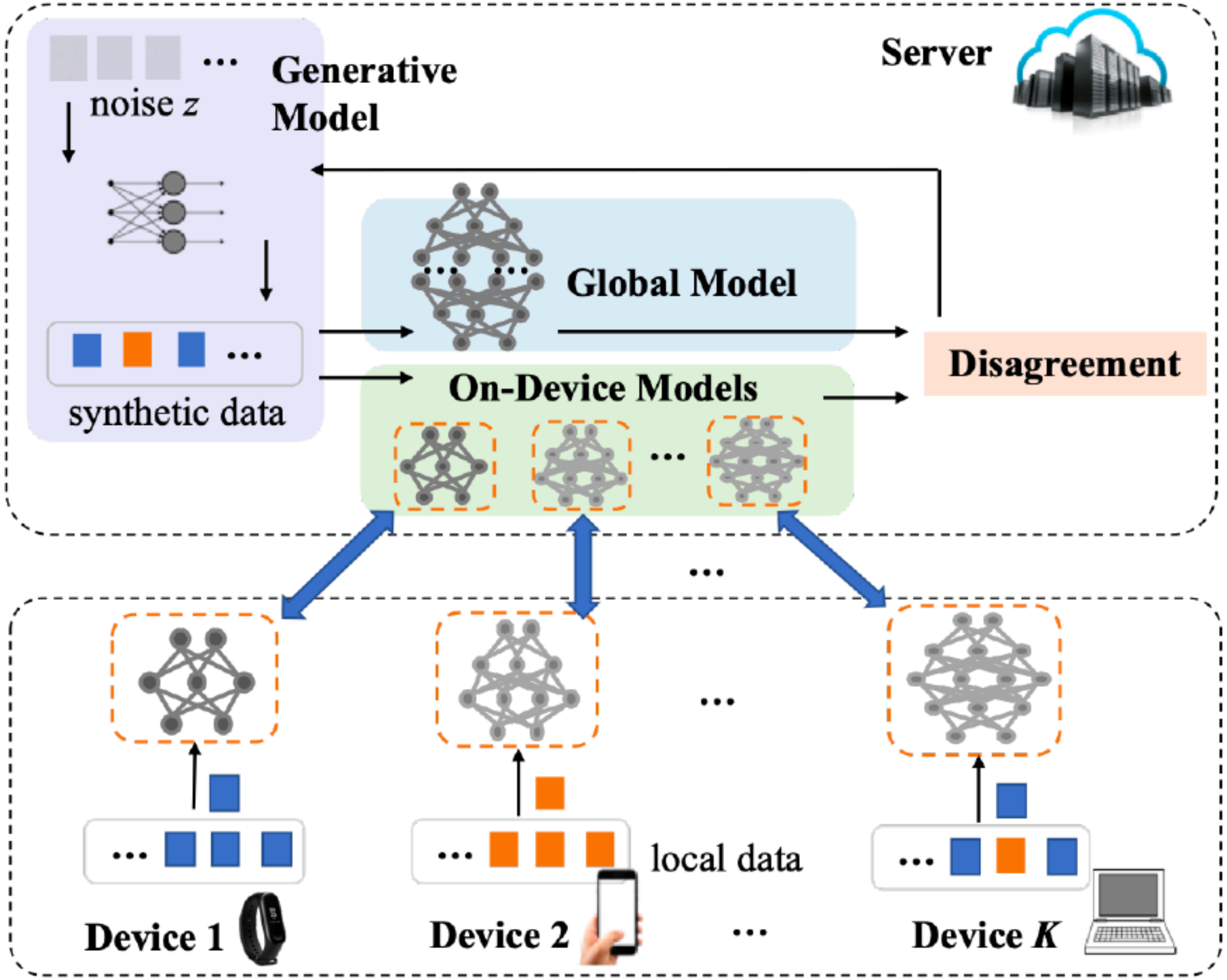
FedZKT

$$\min_{w_k^t} \sum_{\{x,y\}\in\mathcal{D}_k} \mathcal{L}_{CE}(f_k(x; w_k^t), y) + ||w_k^t - w_k^{t-1}||_2^2,$$

$\ell 2$ Regularization (proximal operator) for Non-IID Data Distribution

- handle data heterogeneity
- limit the update of on-device models when training on their local datasets

FedZKT


Server
...
noise z
Generative
Model
Global Model
synthetic data
On-Device Models
Disagreement
local data
Device 1
Device 2
Device K

Experimental Evaluation

Dataset

- Four widely used image datasets: MNIST, KMNIST, FASHIONMNIST, and CIFAR-10

Model heterogeneity

- Five different neural network architectures for each dataset
- MNIST, KMNIST, and FASHIONMNIST (FASHION) (small)
  - a CNN model, a Fully-Connected Model, and three LeNet-like models with different channel sizes and numbers of layers
- CIFAR-10
  - Two ShuffleNetV2 models, two MobileNetV2 models, and a LeNet-like model

Experimental Evaluation

Federated Learning Settings

- Device number: $K \in \{5, 10, 15, 20\}$ (by default $k=10$)
- Communication rounds

  MNIST, KMNIST, FASHION: $T = 50$, 5 local epochs

  CIFAR-10: $T = 100$, 10 local epochs
- Zero-shot knowledge distillation

  MNIST, KMNIST, FASHION: $n_G = n_S = 200$ iterations

  CIFAR-10: $n_G = n_S = 500$ iterations

  Batch size: 256

  Learning rate: reduced by 0.3 at the half and 3/4 of the total iterations

Experimental Evaluation

Data heterogeneity

1) quantity-based label imbalance
2) distribution-based label imbalance

Baseline approach: FedMD

One most representative data-dependent FL algorithm (public dataset) for heterogeneous on-device models

- MNIST, KMNIST, FASHION: FASHION, MNIST, and FASHION, respectively
- CIFAR-10: CIFAR-100 and SVHN

D. Li and J. Wang, “Fedmd: Heterogenous federated learning via model distillation,” NIPS, 2019.

Accuracy under IID

On-Device Dataset	FedMD		FedZKT
	Public Dataset	Average Accuracy	Average Accuracy
MNIST	FASHION	96.69%	**97.76%**
FASHION	MNIST	**85.83%**	84.42%
KMNIST	FASHION	84.02%	**86.43%**
CIFAR-10	CIFAR-100	67.34%	**78.02%**
CIFAR-10	SVHN	20.38%	

TABLE I: Performance of FedZKT and FedMD under IID on-device data distribution.

- The performance of FedMD depends on the selection of the public dataset.

Learning curves under IID

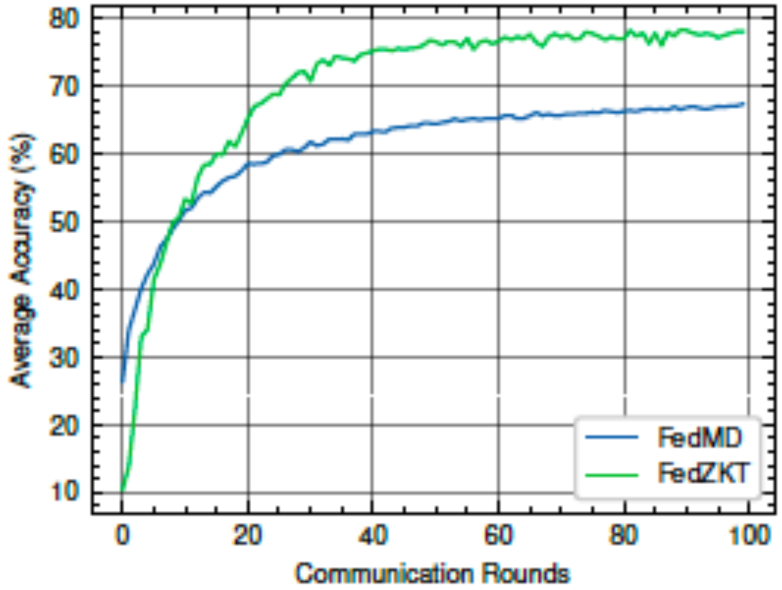


Fig. 3: Learning curves of FedZKT and FedMD (CIFAR-10, IID).

- FedZKT can iteratively produce more representative samples.

Accuracy under Non-IID

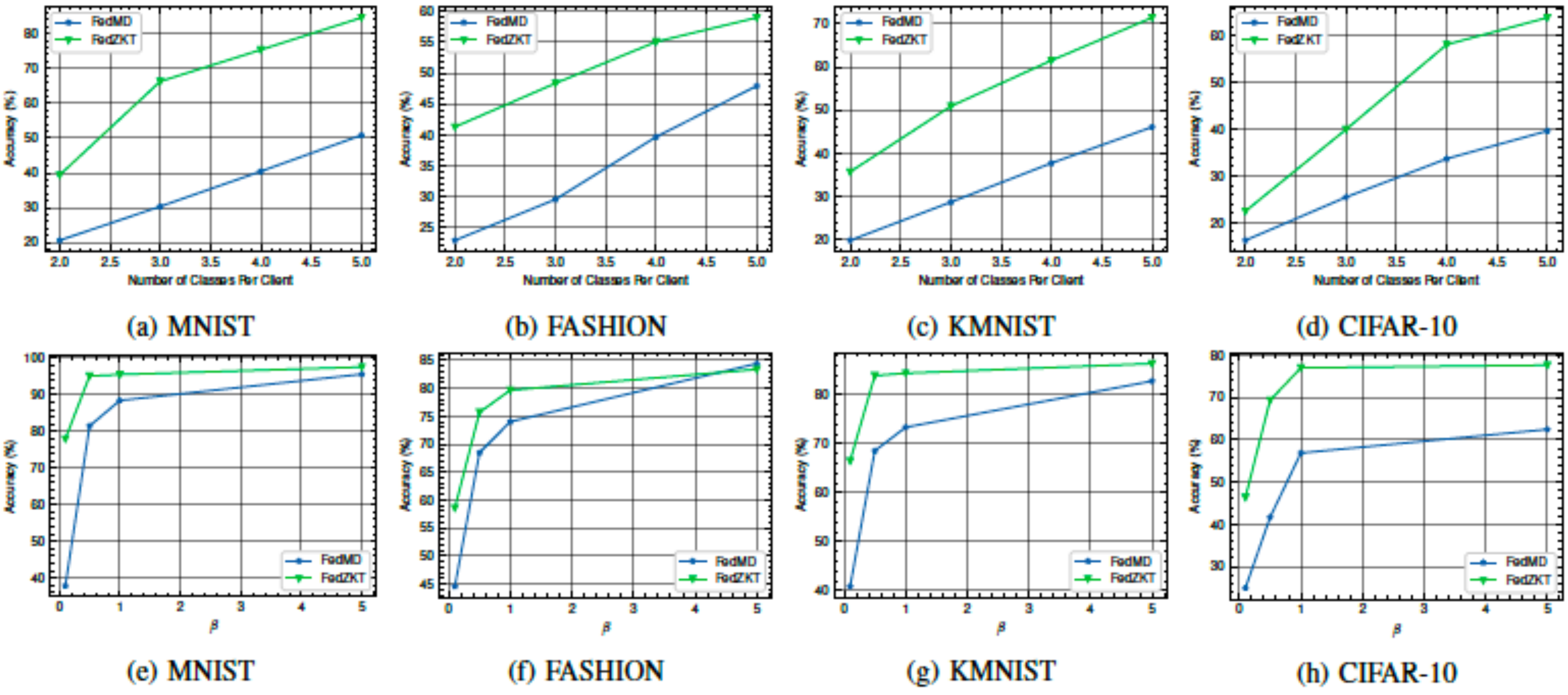


Fig. 4: Performance of FedZKT and FedMD under non-IID on-device data distribution: Quantity-based label imbalance (a)-(d), Distribution-based label imbalance (e)-(h).

- Robustness of FedZKT

Ablation study

Effects of loss functions

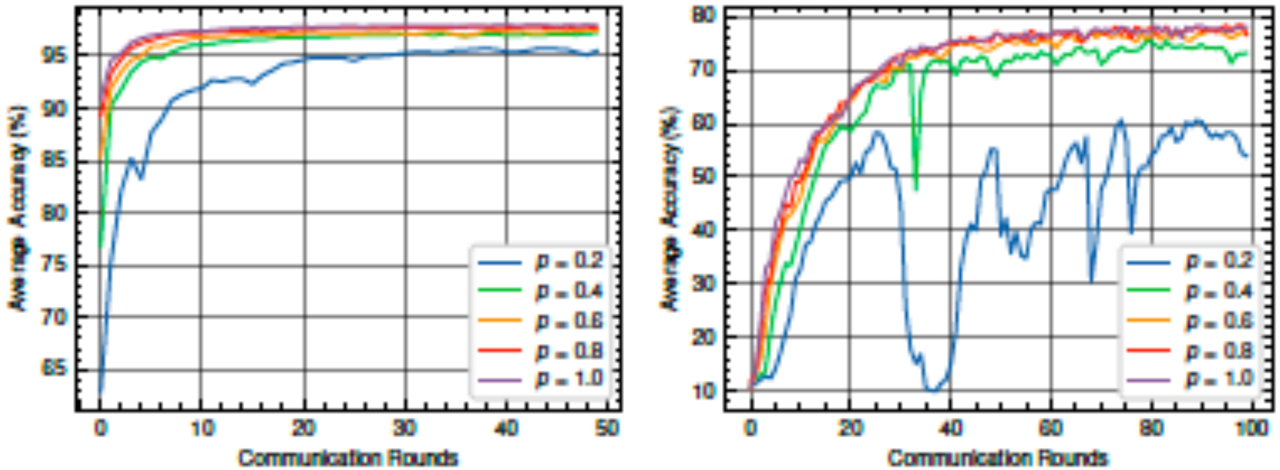
Non-IID scenario	KL-divergence	ℓ_1 norm	SL loss
$C = 5$	48.23%	14.60%	**63.89%**
$\beta = 0.5$	66.17%	16.34%	**69.39%**

TABLE II: Effect of loss functions for zero-shot knowledge distillation in FedZKT (CIFAR-10, Non-IID).

- $\ell 1$ norm loss is not suitable for zero-shot federated distillation under non-iid settings due to the unstable learning performance, although it can avoid the gradient vanishing in zero-shot distillation.

Ablation study

Straggler effect



(a) MNIST, IID. (b) CIFAR-10, IID.

Fig. 6: Effect of stragglers: average accuracy of FedZKT when p portion of devices are trained in each round.

- a portion p of devices as the active ones, $p \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$

Ablation study

Effects of $\ell 2$ Regularization

Non-IID scenario	no regularization	ℓ_2 regularization
$C = 5$	56.58%	**63.89%**
$\beta = 0.5$	66.17%	**69.39%**

TABLE IV: Effect of ℓ_2 regularization in FedZKT (CIFAR-10, Non-IID).

Conclusion

- Propose an innovative FL framework, FedZKT, for resource-constrained and heterogeneous devices in a data-free manner.
- Allow independent on-device model design
- Enable knowledge transfer across heterogeneous on-device models devices via zero-shot knowledge transfer with SL loss function.
- Assign the compute-intensive distillation task to the server to meet the imbalanced capability between server and devices.
- Demonstrate the effectiveness and the robustness of FedZKT through extensive experiments.

Thank You!

Questions?