nature communications

Article

https://doi.org/10.1038/s41467-023-38794-x

Differentially private knowledge transfer for federated learning

Received: 27 April 2023

Accepted: 15 May 2023

Published online: 24 June 2023

Check for updates

Tao Qi[1], Fangzhao Wu[2] ✉, Chuhan Wu[1] ✉, Liang He[1], Yongfeng Huang[1,3,4] ✉ & Xing Xie[2]

Extracting useful knowledge from big data is important for machine learning. When data is privacy-sensitive and cannot be directly collected, federated learning is a promising option that extracts knowledge from decentralized data by learning and exchanging model parameters, rather than raw data. However, model parameters may encode not only non-private knowledge but also private information of local data, thereby transferring knowledge via model parameters is not privacy-secure. Here, we present a knowledge transfer method named PrivateKT, which uses actively selected small public data to transfer high-quality knowledge in federated learning with privacy guarantees. We verify PrivateKT on three different datasets, and results show that PrivateKT can maximally reduce 84% of the performance gap between centralized learning and existing federated learning methods under strict differential privacy restrictions. PrivateKT provides a potential direction to effective and privacy-preserving knowledge transfer in machine intelligent systems.

In recent years, machine learning technology has developed rapidly and empowered intelligent systems in many real-world scenarios, such as intelligent healthcare[1–3] and social computing[4–7]. The success of machine learning usually lies in summarizing useful knowledge from big data, which is mainly benefited from the high capacity and complexity of models[8–10]. Learning machine learning models on centralized data is a mainstream knowledge transfer paradigm[11]. However, training data in many tasks are highly privacy-sensitive[12,13], and recent privacy leakage accidents have drawn more and more attention from the public to the security of user privacy[14–16]. Moreover, some strict privacy regulations such as GDPR[17] and CCPA[18] are also spawned to limit the collection, processing, and storage of user data[19–21]. Due to the privacy-sensitive nature of training data, the centralized model training usually arouses serious privacy concerns and even violates privacy regulations[22–24].

Federated learning (FL) can transfer knowledge from decentralized data[8,25,26], and thereby begins to serve as a privacy-aware model training framework in many privacy-sensitive applications, such as Covid-19 patient detection[27,28] and intelligent personal assistant[29,30]. In federated learning, knowledge is usually extracted from decentralized data into local model updates, and further aggregated into a shared model by communicating local model updates rather than raw data (Fig. 1a)[8,24,31–33]. However, model updates usually have enough capacity to memorize the private information in training data, the disclosure of which still has the risks of leaking raw data[34,35]. To improve privacy security, some federated learning methods propose to transfer knowledge based on a large-scale unlabeled public dataset (Fig. 1b)[36–38]. In these methods, each client first extracts knowledge from decentralized data into local predictions on the entire unlabeled public dataset, then the server aggregates the uploaded local predictions to update a global model. However, the local model predictions are also correlated to private information in local data, and thereby user privacy is still not guaranteed in these works[39].

To protect user privacy in federated learning, local differential privacy (LDP) that perturbs communicated data with noise can be applied to offer theoretical privacy guarantees[40]. Nevertheless, the

[1]Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China. [2]Microsoft Research Asia, 100080 Beijing, China. [3]Zhongguancun Laboratory, 100094 Beijing, China. [4]Institute for Precision Medicine, Tsinghua University, 102218 Beijing, China. ✉e-mail: fangzwu@microsoft.com; wuchuhan15@gmail.com; yfhuang@tsinghua.edu.cn

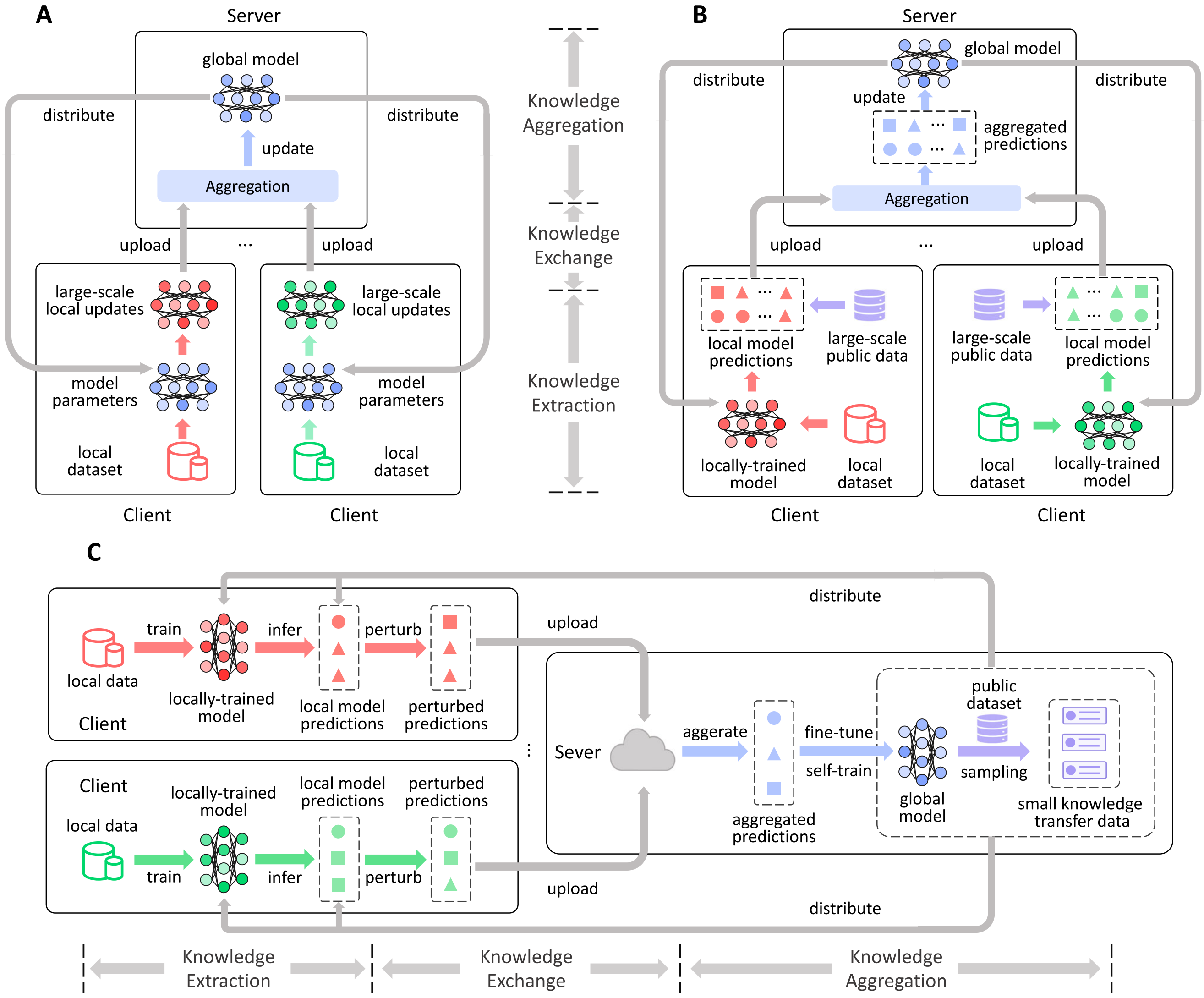

Fig. 1 | The differences between our work and prior methods. A The federated learning method that transfers knowledge via local model updates. **B** The federated learning method that transfers knowledge via local model predictions on a large-scale unlabeled public dataset. **C** The proposed PrivateKT method that transfers knowledge via actively selected small public data.

intensity of LDP noise is usually proportional to the size of communicated data[41], making it ineffective for LDP to protect existing federated learning methods that depend on large-scale public data for knowledge transfer. A naive solution is to reduce the size of public data, however, knowledge transfer via small data is usually ineffective and may seriously degrade the model performance[42]. Thus, in this work, we study how to effectively and privately transfer knowledge via a small amount of data. The core idea of our solution is, that the server actively selects a small amount of public data informative for model learning as the carrier of knowledge transfer (Fig. 1c; named as PrivateKT). In order to achieve privacy guarantees, PrivateKT locally perturbs predictions of local models on the carefully selected data via randomized response mechanism[43]. Then the server in PrivateKT collects and aggregates the local knowledge to update the model. We evaluate PrivateKT on three benchmark datasets for three different real-world tasks. Extensive experiments show that, under strict privacy guarantees, many existing federated learning methods almost degrade into random guesses while PrivateKT can achieve comparable performance with centralized learning (1% loss minimally). Through extensive analysis, we also reveal that small carefully selected data has enough information capacity to transfer big knowledge, which can inspire researchers to design private, effective, and efficient knowledge transfer systems in the future.

Results

Overall framework

Next, we first briefly introduce the overall framework of PrivateKT for private knowledge transfer (Fig. 1c). It can extract the local knowledge from decentralized data to collaboratively learn an intelligent model under differential privacy guarantees. In PrivateKT, training data is locally kept by different clients and never shared with the outside, and an intelligent model is maintained by a central server and has a local copy on each client. Besides, following previous works[36,37], we assume that there is an unlabeled public dataset that is non-privacy sensitive and can be shared across different parties for knowledge transfer.

The knowledge transfer in PrivateKT includes three steps, namely, knowledge extraction, knowledge exchange, and knowledge aggregation. In the knowledge extraction step, each client first trains the intelligent model on its local data and then computes model predictions on a small amount of knowledge transfer data (named KT data). The KT data is actively sampled from the public dataset by the server based on an importance sampling mechanism, where the public data with lower model confidence will be assigned a higher sampling opportunity. In the knowledge exchange step, we share the local knowledge with the server under differential privacy guarantees. Each client locally perturbs model predictions via the randomized response mechanism[43], which randomly chooses whether to replace a local model prediction with a randomly-generated category label before sending it to the server. In the knowledge aggregation step, the server aggregates uploaded perturbed local predictions and stores them in a knowledge buffer. In this way, the historical aggregated knowledge in the past several rounds is kept by the knowledge buffer, and can be further encoded into the global model by fine-tuning the global model on the knowledge buffer. Moreover, we employ a self-training method[44] to update the global model on the unlabeled public data to accelerate the model convergence. By repeating this process, we can privately transfer high-quality knowledge from decentralized data into the shared intelligent model.

Performance evaluation

We conduct experiments in three real-world tasks for performance evaluation. The first task is handwritten digit classification, which needs to classify the category of a digit written by a user. It is based on a widely used federated learning benchmark dataset (named MNIST)[45]. The second task is text-based disease prediction, which needs to predict the diseases described in medical abstracts. It is based on a public dataset (named MedText) released by a Kaggle medical text mining competition. The third task is image-based pneumonia detection, which aims to detect pneumonia from chest X-ray images. This task is based on a real-world dataset (named X-ray) released in a Kaggle X-Ray analysis competition. Following the previous work[36], 20% of training data is used as the candidate unlabeled data pool to choose samples for knowledge transfer. The model performance is compared under both independent identical data distribution (IID) and non-identical independent data distribution (Non-IID). The non-IID data distributions used for evaluation in experiments include the class non-IID distribution where data classes of local clients are imbalanced, the size non-IID distribution where data sizes of local clients are imbalanced, and the mixed non-IID distribution where both data classes and data sizes of local clients are imbalanced. The IID and Non-IID data partition strategies follow the settings of previous work[46]. (We show more details of datasets in the Supplementary Information.)

The architecture of the basic models trained on MNIST and X-Ray is a two-layer convolution network[45], and the basic model trained on MedText is a transformer network[47]. Several representative knowledge transfer methods for federated learning methods are compared in experiments, including FedSGD[8], FedAvg[8], FedAdam[31], FedMD[37], and FedED[36]. The comparison includes results of using LDP or not. We apply LDP to protect the baseline federated learning methods by adding noise to the exchanged local model updates or model predictions. We use the definition of ϵ-LDP and the privacy budget ϵ is set to 5 (see the "Methods" section). We also include the results of centralized model training (named CenTrain) as a reference for performance comparisons. We use the accuracy as the metric for evaluation on MNIST, and the Macro-F1 as the metric for evaluation on MedText and X-Ray. (More detailed experiential settings are presented in the Supplementary Information.)

We independently repeat each experiment five times and report the average performance with standard deviations (Fig. 2). From the results, we find that without differential privacy both PrivateKT and other federated learning methods achieve comparable performance with centralized model training. However, when we apply LDP to protect user privacy, the LDP noise seriously hurts the performance of existing federated learning methods. For example, FedAvg drops 86.85% of accuracy on MNIST under the IID data distribution, and FedMD drops 41.91% of Macro-F1 on X-Ray under the class non-IID data distribution. This is because these federated learning methods transfer knowledge by exchanging a large volume of intermediate variables (such as local model updates). However, the intensity of LDP noise is usually proportional to the size of communicated data, making it ineffective for these methods to balance knowledge utility and privacy protection. By contrast, PrivateKT effectively improves the performance of federated learning under the same differential privacy guarantees, and significantly outperforms previous methods ($p < 1e-4$ based on t-test). This is because PrivateKT uses small carefully selected data to condense high-quality knowledge, which does not substantially suffer from the perturbation of LDP noise meanwhile improving the effectiveness of knowledge transfer. We then analyze the contributions of different mechanisms in PrivateKT in the following section. We also compare the efficiency and generality of PrivateKT with other FL methods in Supplementary Information. It is worth noting that the evaluation presented in Fig. 2 is primarily founded on moderately large datasets with thousands or tens of thousands of samples (e.g., MNIST and X-Ray). Therefore, further exploration is necessary to fully assess the effectiveness of PrivateKT on large-scale datasets. In light of this, we undertake further analysis by comparing various methods on two larger FL benchmark datasets, CIFAR-10 and CIFAR-100[48], which are expounded upon in the Supplementary Information. Results show that the main conclusions of our paper still hold: PrivateKT effectively improves the performance of other FL methods under strict privacy restrictions.

Model effectiveness

Next, we verify the impacts of several important mechanisms in PrivateKT on knowledge transfer, i.e., knowledge buffer, importance sampling, and self-training (Fig. 3). We remove the knowledge buffer and the self-training method from PrivateKT, and replace the importance sampling mechanism with a uniform sampling method, individually, to verify their effectiveness. Results show that both the knowledge buffer and the importance sampling mechanisms effectively improve the model performance and accelerate the model convergence. This is because, in order to mitigate the damage of LDP noise on model performance, PrivateKT only uses a small amount of data sampled from an unlabeled public dataset for knowledge transfer. However, small public data may be insufficient for the effective knowledge transfer and result in suboptimal model performance. Thus, in PrivateKT we tackle this challenge from two aspects. First, we propose to sample knowledge transfer data based on their informativeness for model training, to maximize the quality of knowledge carried by the small sampled data. Second, we propose a knowledge buffer to store and encode historical useful knowledge to the global model, aiming to incorporate more useful knowledge for model updating. Moreover, to enhance the knowledge transfer in PrivateKT, we also employ the self-training technique to further fine-tune the global model, whose contribution is also verified by the results. These results show that PrivateKT can exploit small data for transferring big knowledge. This finding reveals that, big knowledge is not necessarily obtained from big data, but also can be mined from small but representative data.

Next, we analyze the trade-off between the performance and privacy of the knowledge transfer in PrivateKT (Fig. 4). We show the performance of PrivateKT under different privacy budgets (denoted as ϵ) and different knowledge transfer sample sizes (denoted as K), where a smaller privacy budget means a stronger differential privacy guarantee. We find that strong differential privacy guarantees do not seriously hurt the model accuracy. For example, the best accuracy of PrivateKT on MNIST is around 94% under a strong privacy guarantee,

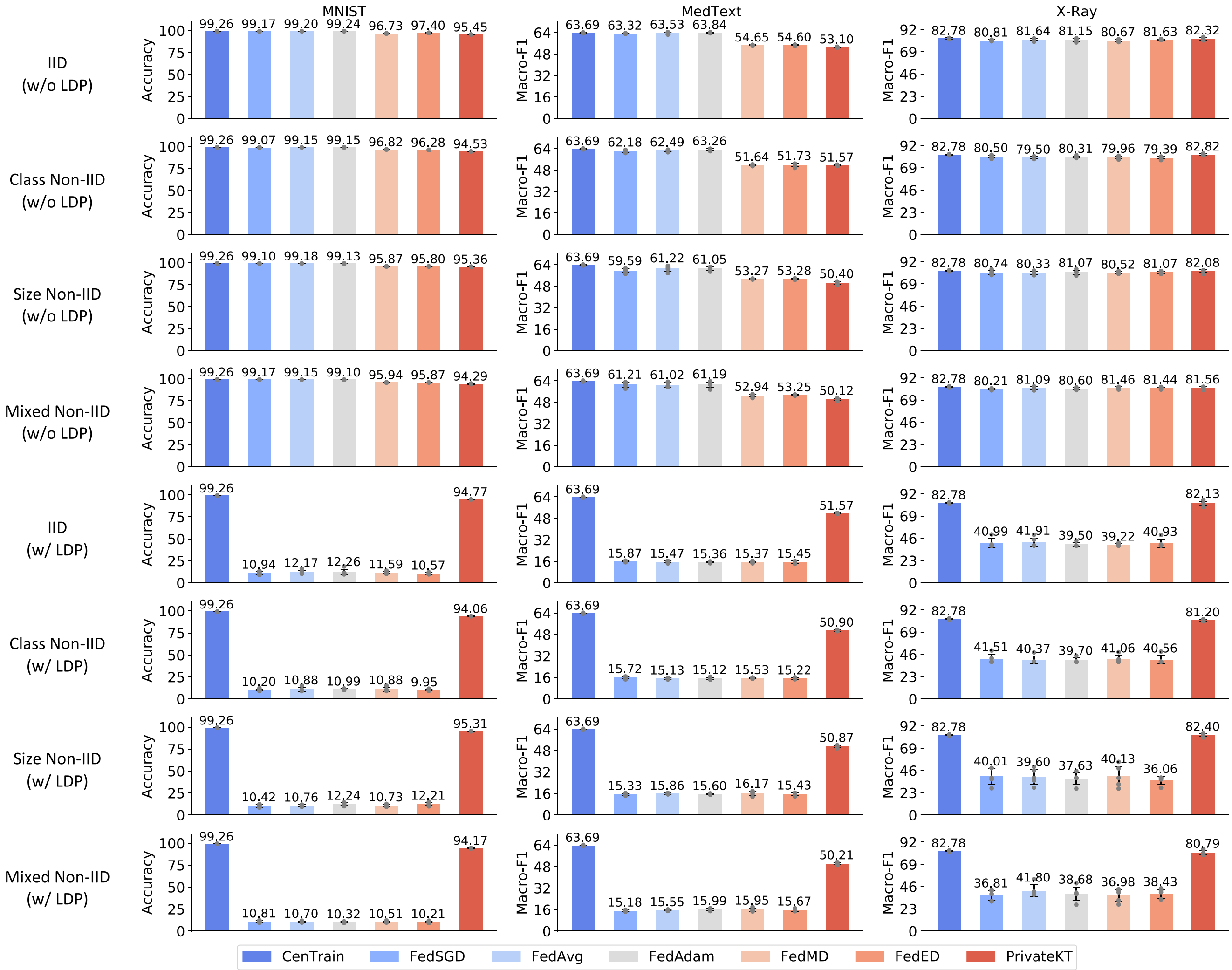

Fig. 2 | Model performance of different methods on three datasets. The error bars represent the mean results with standard deviations. The comparisons are based on different data distributions and using LDP or not ($\epsilon = 5$). The data distributions includes the IID data distribution, imbalanced data class distribution (Class Non-IID), imbalanced data size distribution (Size Non-IID), and distribution where both local data sizes and classes are imbalanced (Mixed Non-IID). The averaged performance and corresponding standard deviations are shown. The results show that the LDP noise seriously hurts the performance of existing federated methods, and PrivateKT can significantly improve the model performance under the same privacy guarantees ($p < 1e-4$ based on t-test).

i.e., $\epsilon = 2$. The results verify that PrivateKT can effectively balance the knowledge utility and privacy protection in federated learning. We also find that in each round of PrivateKT a small number of samples (e.g., 2) are sufficient for privately and effectively extracting knowledge from decentralized data to a shared model, which further confirms that, small data is possible for transferring big knowledge. More detailed analyses on the hyper-parameter settings of PrivateKT are shown in Supplementary Information.

Discussion

From big data to big knowledge, is an important vision of the current machine learning research[8,10]. Federated learning that transfers knowledge from decentralized data into a shared intelligent model is widely used to reduce user privacy risks during knowledge transfer[8,23]. Nevertheless, its privacy security is not guaranteed and needs to be protected by some privacy protection methods such as local differential privacy[40]. However, in this work, we discover that previous federated learning methods are less performant in trading off privacy protection and knowledge utility. Though simply reducing the size of exchanged data is a potential solution to this problem, it may substantially degrade the knowledge transfer effectiveness and thereby yield suboptimal model accuracy[42]. Thus, there raises a question that whether it is possible to transfer big knowledge via small data with strong privacy guarantees.

In this work, we reveal that the answer to the above question is true. The core of achieving this goal is selecting public samples based on their informativeness rather than randomly. After knowledge extraction through these representative samples on local clients, high-quality knowledge encoded by local models can be transferred to the server for aggregation to prepare for the further round of updates. This paradigm enables privacy-preserving knowledge transfer on small data that can minimally suffer from the performance degradation brought by differential privacy, which shows a novel direction to train machine learning models on decentralized data to exploit swarm intelligence. It can also attract further attention to more sophisticated data exploitation methods, rather than simply collecting and involving more and more training data, which is not beneficial for understanding the bound of machine intelligence under limited real-world data and reducing the environmental pollution brought by the computation. We hope our work can further inspire researchers to facilitate knowledge engineering in a more effective, efficient, and privacy-preserving way.

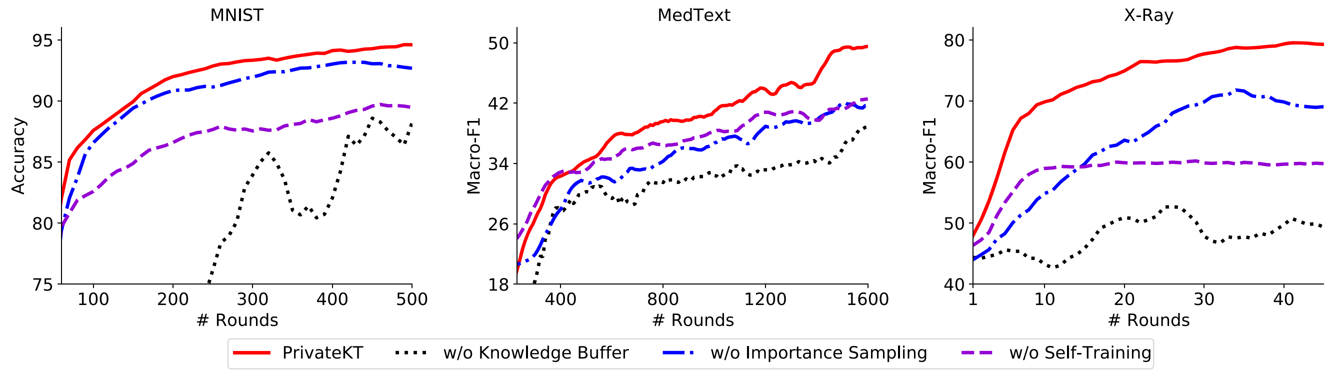

Fig. 3 | The impacts of important mechanisms in PrivateKT on knowledge transfer. We analyze the impacts of the knowledge buffer and the self-training mechanisms by removing them from PrivateKT individually, and the impact of the importance sampling mechanism by replacing it with a uniform sampling method. Results show that all of these mechanisms can enhance knowledge transfer in PrivateKT, and the knowledge buffer makes the greatest contribution.

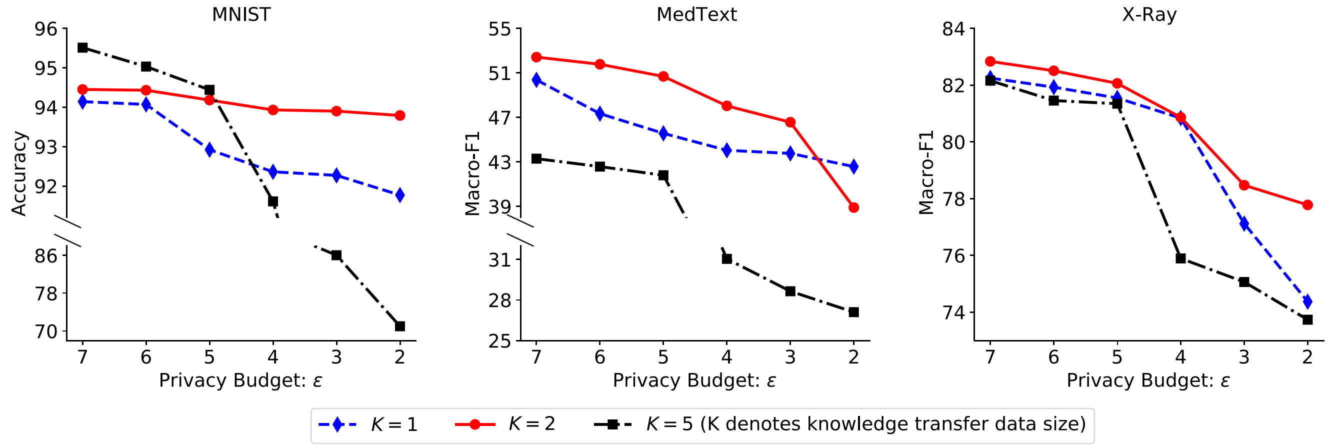

Fig. 4 | Analysis on the utility-privacy tradeoff in PrivateKT. The model performance under various privacy budgets (i.e., ϵ) and sizes of knowledge transfer samples (i.e., K) are presented in this figure. The results show that PrivateKT can achieve satisfactory performance under very strong privacy guarantees (e.g., $\epsilon = 2$).

However, our work also has the following limitations. First, the private knowledge transfer method in PrivateKT requires an unlabeled dataset that can be shared across different parties, which may be inaccessible in some applications (e.g., personalized e-commerce). Fortunately, some latest research works find that the knowledge transfer can be effectively performed in a data-free manner[49–51]. Thus, we plan to apply the data-free knowledge transfer methods to PrivateKT to improve its practicability in real-world scenarios. Second, in real-world applications, PrivateKT has the risk of being attacked by Byzantine clients[52]. Therefore, in our future work, we plan to study how to defend the attack to improve the robustness of PrivateKT based on previous works[52–54]. Third, compared with the privacy-invasive centralized training, PrivateKT faces more significant performance degradation on larger datasets. This is because, according to the no free lunch theorem for privacy security and algorithm utility of federated learning[55], stronger privacy protection will lead to poorer algorithm performance. Since preserving the privacy of a larger volume of training data usually needs stronger protection, the performance degradation of PrivateKT will also become more serious. Thus, in our future work, we will explore improving the privacy-preserving knowledge transfer mechanism of PrivateKT to approach the theoretical performance upper bound. Besides, the demonstration of PrivateKT is mainly based on moderately large datasets (e.g., MNIST and CIFAR), and we admit that the superiority of PrivateKT over other FL methods is not guaranteed on large-scale datasets. In practice, we think our approach is applicable in most scenarios with small or moderate data volumes, and the scalability on huge datasets needs further exploration.

Methods

Next, we will present the differentially private knowledge transfer method for federated learning (named PrivateKT). We will first give former definitions of local differential privacy and the research problem studied in this paper, and then introduce the details of our PrivateKT method.

Preliminary

The local differential privacy method (LDP)[56] aims to protect user privacy under theoretical guarantees. The core idea of LDP is to perturb the shared data via a randomized mechanism to guarantee privacy security. Formally, the definition of LDP can be summarized as follow: a randomized mechanism $\mathcal{M}(\cdot)$ can protect the input data · under ϵ-LDP, if and only if for two arbitrary input data X and X', and any output $Y \in range(\mathcal{M})$, the following inequation holds:

$$Pr[\mathcal{M}(X) = Y] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(X') = Y], \tag{1}$$

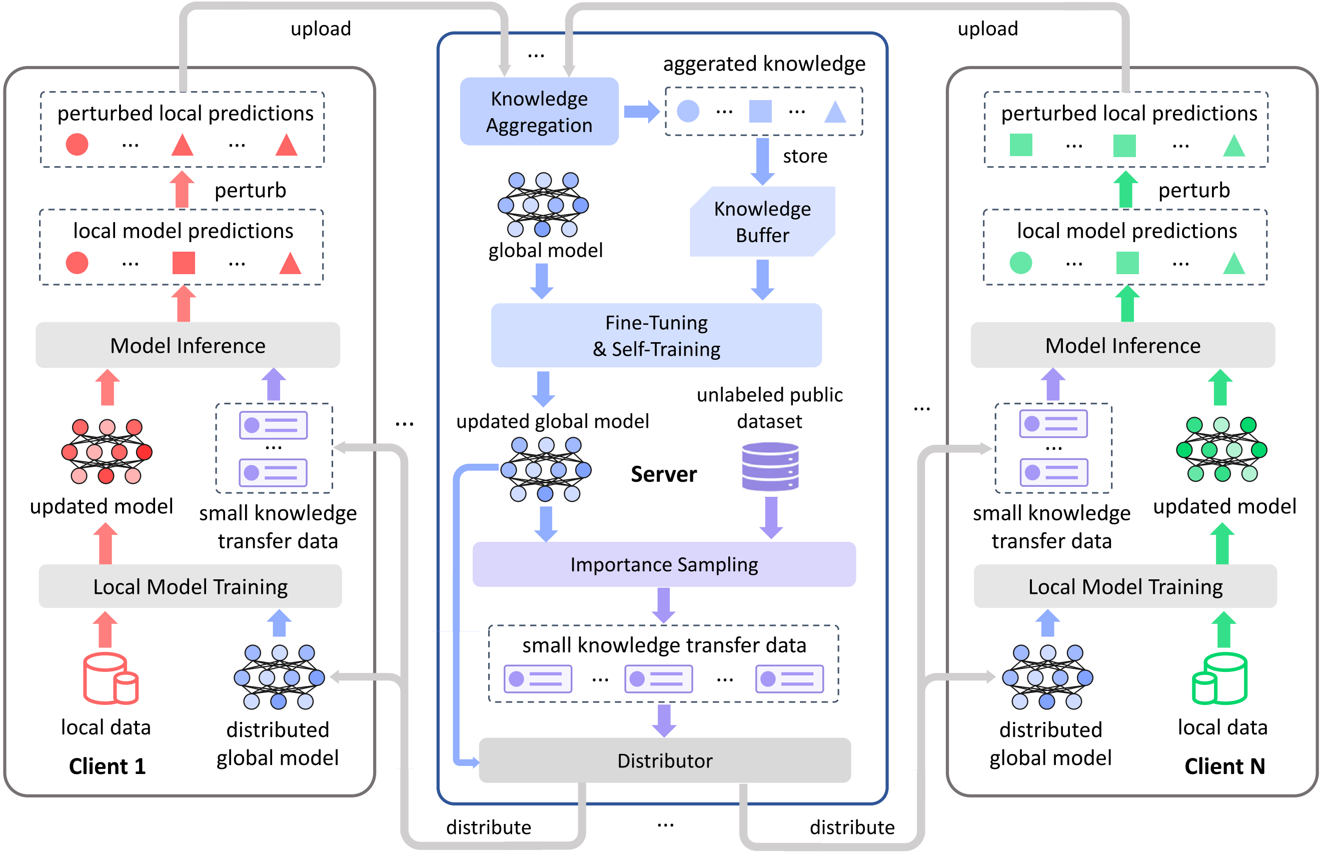

Fig. 5 | The framework of our PrivateKT method.

where $Pr[\cdot]$ is the probability of $\cdot$, and ϵ is the privacy budget. The privacy budget ϵ quantifies the privacy guarantee, where a smaller privacy budget means stronger privacy protection.

Problem definition

Following popular federated learning settings, PrivateKT includes N clients and a central server. Each client privately keeps its local dataset and never shares it with the outside, where the local dataset in the i-th client is denoted as $\mathcal{D}_l^i$. The global model is maintained by the central server and has a local copy on each client. The central server is also responsible for coordinating the clients to participate in the knowledge transfer. In addition, we assume that there is an unlabeled public dataset $\mathcal{D}_p$ that is non-privacy sensitive and can be shared across different parities for knowledge transfer, where the i-th sample in $\mathcal{D}_p$ is denoted as x_i^p. In order to guarantee privacy security during knowledge transfer, any communicated variables correlated to the local private data need to be protected by the LDP method. The research problem studied in this paper is to design a both private and effective knowledge transfer method for federated learning.

Differential private knowledge transfer

The core of private knowledge transfer is communicating perturbed local model predictions on a small amount of actively selected public data. By drastically reducing the size of communicated variables, PrivateKT can effectively mitigate the damage of LDP noise on model performance. Nevertheless, randomly sampled small data may be insufficient to transfer high-quality knowledge from local data to a global model. Thus, we further propose several mechanisms to improve the effectiveness of knowledge transfer based on small data. Next, we will introduce the details of the differential private knowledge transfer in PrivateKT (Fig. 5).

Take the t-th knowledge transfer round as an example, PrivateKT includes three core steps, i.e., knowledge extraction, knowledge exchange, and knowledge aggregation. The knowledge extraction step aims to extract knowledge from local data and encode it into local predictions on small actively sampled data. Specifically, the server first distributes the global model in the t-round (denoted as Θ_t) and K pieces of knowledge transfer (KT) data to each client, and selects a part of clients for model training, where the selected client set is denoted as $\mathcal{G}_t$. (The sampling mechanism of KT data will be introduced in the next paragraph.) For an arbitrary client $c \in \mathcal{G}_t$, it first trains the latest model Θ_t on its local dataset $\mathcal{D}_l^c$. Then the client c computes predictions of the locally-trained model on the KT data for knowledge extraction, where x_i^t denotes the i-th KT data and $y_{c,i}^t$ denotes the local model prediction of the client c on x_i^t. In this way, knowledge can be extracted from local data into local model predictions, and exchanging local predictions can transfer local knowledge to the central server.

However, the local model predictions are correlated to the private data, the disclosure of which remains the risk of leaking raw data. Thus, to guarantee user privacy security under LDP, each client locally perturbs local predictions via the randomized response mechanism[43]. Specifically, for each local model prediction $\mathbf{y}$, each client c randomly chooses whether replace it with a randomly-generated category label $\mathbf{f}$ before uploading it to the server:

$$\hat{\mathbf{y}} = \begin{cases} \mathbf{y}, & R=1 \\ \mathbf{f}, & R=0 \end{cases}, \quad R \sim \mathcal{B}(\beta), \quad \mathbf{f} \sim \mathcal{P}(C), \tag{2}$$

where $\mathbf{y} \in \{0,1\}^C$ is the one-hot category vector predicted by the local model, $\mathbf{f} \in \{0,1\}^C$ is a random one-hot vector drawn from a uniform multinomial distribution $\mathcal{P}(C)$, $\hat{\mathbf{y}}$ is the perturbed local prediction, R is a random variable drawn from a Bernoulli distribution $\mathcal{B}(\beta)$, C is the number of classification categories and β is the probability of assigning the Bernoulli random variable R to 1. Based on the randomized response mechanism, the client c can build the perturbed local predictions $\{\hat{\mathbf{y}}_{c,i}^t | i = 1,2,...,K\}$ for the knowledge transfer data. By uploading the perturbed predictions to the server, we can privately exchange local knowledge under differential privacy guarantees. (Discussions on privacy guarantees are in the next section.)

After the server collects perturbed predictions from selected clients $\mathcal{G}_t$, the knowledge aggregation step can be executed to update the global model. The sever first aggregates the local predictions on the same KT data to estimate the averaged predictions of different local models on it. Take the i-th knowledge transfer data x_i^t as an example, the averaged prediction $\mathbf{y}_i^t = \frac{1}{|\mathcal{G}_t|}\sum_{c\in\mathcal{G}_t}\mathbf{y}_{c,i}^t$ on x_i^t is estimated based on the following equation:

$$\hat{\mathbf{y}}_i^t = \frac{1}{\beta}\left(\frac{1}{|\mathcal{G}_t|}\sum_{c\in\mathcal{G}_t}\hat{\mathbf{y}}_{c,i}^t - \frac{1-\beta}{C}\mathbf{1}\right), \qquad (3)$$

where $\hat{\mathbf{y}}_i^t$ is an unbiased estimation of $\mathbf{y}_i^t$ and the mean square error of the estimation can asymptotically converge to 0. (The proof is in Supplementary Information.) In this way, the LDP noise can be reduced in the aggregated knowledge, and fine-tuning the global model on the aggregated knowledge can effectively mitigate the damage of LDP noise on knowledge transfer.

Recall that, due to the proportional relation between the LDP noise intensity and communicated data volume, in PrivateKT only a small amount of public data is used for knowledge transfer to mitigate the damage of LDP noise. However, small data may be insufficient to serve as a high-quality carrier to transfer knowledge, which may lead to a suboptimal model performance. To tackle this challenge, we propose two mechanisms to enhance knowledge transfer from different aspects. First, we propose an importance sampling mechanism to maximize the knowledge capacity of KT data for training the global model Θ^t. In this mechanism, we measure the uncertainty of the global model Θ^t on each unlabeled data in $\mathcal{D}_p$ based on the information entropy, and assign a higher sampling opportunity to unlabeled data with higher model uncertainty. The model uncertainty u_i^d and the sampling weight w_i^d of the i-th unlabeled data x_i^p in $\mathcal{D}_p$ are computed as follow:

$$w_i^d = \frac{\exp(u_i^d)}{\sum_{j=1}^{|\mathcal{D}_p|}\exp(u_j^d)}, \quad u_i^d = -\sum_{j=1}^{C} p(x_i^p, j; \Theta_t)\log p(x_i^p, j; \Theta_t), \qquad (4)$$

where $p(x_i^p, j; \Theta_t)$ is the probability of classifying x_i^p to the j-th category based on model Θ_t. Second, we propose a knowledge buffer to store historical aggregated knowledge, aiming to encode more useful knowledge to the global model. The server first stores the aggregated knowledge of the current round in the knowledge buffer and then utilizes the knowledge in the buffer to fine-tune the global model Θ_t. (The updated global model is denoted as Θ_t'.) The knowledge buffer is of size B and maintains the stored knowledge in a first-in-first-out manner.

Moreover, to accelerate the model convergence, we employ the self-training technique[44] to further fine-tune the global model Θ_t'. We randomly select M samples with low model uncertainties from $\mathcal{D}_p$ and utilize them to self-train the model Θ_t':

$$w_i^s = \frac{\exp(-u_i^s)}{\sum_{j=1}^{|\mathcal{D}_p|}\exp(-u_j^s)}, \quad u_i^s = -\sum_{j=1}^{C} p(x_i^p, j; \Theta_t')\log p(x_i^p, j; \Theta_t'), \qquad (5)$$

where u_i^s is the uncertainty of model Θ_t' on x_i^p and w_i^s is the sampling opportunity of x_i^p for the self-training. Until now, we have finished a knowledge transfer round in PrivateKT and privately transferred knowledge from decentralized data to the global model, where the updated model is denoted as Θ^{t+1}. Furthermore, we can continue the next knowledge transfer round, after the server distributes the latest global model Θ^{t+1} and corresponding KT data to local clients. By repeating this process, we can transfer knowledge from decentralized data to collaboratively learn an intelligent model in an effective and privacy-preserving way. The workflow of PrivateKT is also summarized in Algorithm 1.

Algorithm pseudo code

Algorithm 1. Workflow of PrivateKT

```
Setting the hyperparameters $\epsilon$, $K$, $\beta$, $B$, $M$ and $T$
Sever randomly initializes the model parameter $\Theta_1$
Server randomly selects $K$ knowledge transfer data $\mathcal{D}_d^1 = \{x_i^1 | i = 1,...,K\}$ from $\mathcal{D}_p$.
for $t$ in $1, 2, ..., T$ do
    Sever distributes $\Theta_t$ and $\mathcal{D}_d^t$ to each client
    Server randomly selects a group of clients $\mathcal{G}_t$
    for each client $c \in \mathcal{G}_t$ (in parallel) do
        Locally train model $\Theta_t$ on the local dataset $\mathcal{D}_l^c$
        for $i$ in $1, 2, ..., K$ do
            Compute local model prediction $\mathbf{y}_{c,i}^t$ on the KT data $x_i^t$
            Randomly draw $R \sim \mathcal{B}(\beta)$ and $\mathbf{f} \sim \mathcal{P}(C)$
            Compute perturbed local model prediction $\hat{\mathbf{y}}_{c,i}^t$ via Eq. (2)
        end for
        Upload perturbed local model predictions to the server
    end for
    Server aggregates local knowledge and stores them in the knowledge buffer of size $B$
    Server fine-tunes the global model $\Theta_t$ on the knowledge buffer
    Server self-trains the global model and builds the updated model $\Theta_{t+1}$
    Server samples knowledge transfer data $\mathcal{D}_d^{t+1}$ via the importance sampling mechanism
end for
```

Discussion on privacy protection

Next, we will discuss the privacy guarantees of the knowledge transfer in PrivateKT. In PrivateKT, the local private data is kept by each client and never shared with the outside. In order to transfer knowledge from decentralized data to an intelligent model, PrivateKT extracts knowledge from local data into predictions on small KT data, and shares them with a central server for knowledge aggregation. Thus, in PrivateKT, among all local variables correlated to the private data, only local predictions are shared with the server. Since the communication of local predictions may leak raw data, we propose to perturb each local prediction before sending it to the central server to protect user privacy. The privacy security of a single knowledge transfer round in PrivateKT is guaranteed by the ϵ-LDP based on Lemma 1. (The proof is in the Supplementary Information.)

Lemma 1. Given the size of knowledge transfer samples (i.e., K), the privacy protection of knowledge transfer in PrivateKT is gauranteed by ϵ-LDP if the following equation holds:

$$\beta = \frac{\exp(\frac{\epsilon}{K}) - 1}{\exp(\frac{\epsilon}{K}) - 1 + C}. \qquad (6)$$

Moreover, in PrivateKT we can further avoid the accumulation of privacy budgets during different knowledge transfer rounds based on the model shuffling method[41,57]. Thus, the privacy security of the whole knowledge transfer process in PrivateKT is also guaranteed by ϵ-LDP, if the condition in Lemma 1 can be satisfied.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets involved in this study are all publicly available ones. The usage of these datasets in this paper is permitted under their licenses. The MNIST dataset is available at http://yann.lecun.com/exdb/mnist/. The MedText dataset is available at https://www.kaggle.com/datasets/chaitanyakck/medical-text. The X-Ray dataset is available at https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia. The CIFAR-10 and CIFAR-100 datasets are available at https://www.cs.toronto.edu/~kriz/cifar.html. The experimental results generated in this study are provided in the Source Data file. Source data are provided with this paper.

Code availability

Codes for this work are available on a public repository[58] https://github.com/taoqi98/PrivateKT. We also provide sufficient details in the Methods and Supplementary Information for implementing experiments in this work.

References

1. Kim, M. et al. Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning. *Nat. Biomed. Eng.* **6**, 267–275 (2022).
2. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
3. Mueller, Y. M. et al. Stratification of hospitalized covid-19 patients into clinical severity progression groups by immuno-phenotyping and machine learning. *Nat. Commun.* **13**, 915 (2022).
4. Aiken, E., Bellue, S., Karlan, D., Udry, C. & Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. *Nature* **603**, 864–867 (2022).
5. Wang, D. et al. Machine learning reveals cryptic dialects that explain mate choice in a songbird. *Nat. Commun.* **13**, 1630 (2022).
6. Heffner, J. & FeldmanHall, O. A probabilistic map of emotional experiences during competitive social interactions. *Nat. Commun.* **13**, 1718 (2022).
7. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. J. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
8. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 1273–1282 (PMLR, 2017).
9. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
10. Kenton, J. D. M.-W. C. & Toutanova, L. K. Bert: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186 (ACL, 2019).
11. Bhagoji, A. N., Chakraborty, S., Mittal, P. & Calo, S. Analyzing federated learning through an adversarial lens. In *ICML*, 634–643 (PMLR, 2019).
12. Han, T. et al. Breaking medical data sharing boundaries by using synthesized radiographs. *Sci. Adv.* **6**, eabb7973 (2020).
13. Wan, Z. et al. Using game theory to thwart multistage privacy intrusions when sharing data. *Sci. Adv.* **7**, eabe9986 (2021).
14. Goroff, D. L. Balancing privacy versus accuracy in research protocols. *Science* **347**, 479–480 (2015).
15. Gibney, E. The scant science behind cambridge analytica's controversial marketing techniques. *Nature* **10** (2018).
16. Gibney, E. Privacy hurdles thwart facebook democracy research. *Nature* **574**, 158–160 (2019).
17. Voigt, P. & Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide*, 1st edn. Vol. 10, p. 3152676 (Springer International Publishing, 2017).
18. Pardau, S. L. The california consumer privacy act: towards a european-style privacy regime in the united states. *J. Technol. Law Policy* **23**, 68 (2018).
19. Price, W. N., Kaminski, M. E., Minssen, T. & Spector-Bagdady, K. Shadow health records meet new data privacy laws. *Science* **363**, 448–450 (2019).
20. Dove, E. S., Chen, J. & Loideain, N. N. Raising standards for global data-sharing. *Science* **371**, 133–134 (2021).
21. Ziegenhain, C. & Sandberg, R. Bamboozle removes genetic variation from human sequence data for open data sharing. *Nat. Commun.* **12**, 6216 (2021).
22. Mohri, M., Sivek, G. & Suresh, A. T. Agnostic federated learning. In *ICML*, 4615–4625 (PMLR, 2019).
23. Eshete, B. Making machine learning trustworthy. *Science* **373**, 743–744 (2021).
24. Li, T. et al. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2**, 429–450 (2020).
25. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
26. Froelicher, D. et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat. Commun.* **12**, 5910 (2021).
27. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nat. Med.* **27**, 1735–1743 (2021).
28. Bai, X. et al. Advancing covid-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat. Mach. Intell.* **3**, 1081–1089 (2021).
29. Yang, T. et al. Applied federated learning: Improving google keyboard query suggestions. *arXiv. preprint arXiv:1812.02903* (2018).
30. Muhammad, K. et al. Fedfast: Going beyond average for faster training of federated recommender systems. In *KDD*, p. 1234–1242 (ACM, 2020).
31. Reddi, S. J. et al. Adaptive federated optimization. In *ICLR* (OpenReview, 2021).
32. Hanzely, F., Hanzely, S., Horváth, S. & Richtárik, P. Lower bounds and optimal algorithms for personalized federated learning. In *NeurIPS*, p. 2304–2315 (Curran Associates Inc., 2020).
33. Wu, C., Wu, F., Lyu, L., Huang, Y. & Xie, X. Communication-efficient federated learning via knowledge distillation. *Nat. Commun.* **13**, 2032 (2022).
34. Zhu, L. & Han, S. *Federated Learning*, p. 17–31 (Springer, 2020).
35. Truex, S., Liu, L., Chow, K.-H., Gursoy, M. E. & Wei, W. Ldp-fed: Federated learning with local differential privacy. In *EdgeSys*, p. 61–66 (ACM, 2020).
36. Sui, D. et al. Feded: Federated learning via ensemble distillation for medical relation extraction. In *EMNLP*, p. 2118–2128 (ACL, 2020).
37. Li, D. & Wang, J. Fedmd: Heterogenous federated learning via model distillation. *arXiv. preprint arXiv:1910.03581* (2019).
38. Lin, T., Kong, L., Stich, S. U. & Jaggi, M. Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*, p. 2351–2363 (Curran Associates Inc., 2020).
39. Sun, L. & Lyu, L. Federated model distillation with noise-free differential privacy. In *IJCAI*, p. 1563–1570 (IJCAI, 2021).

40. Kenny, C. T. et al. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Sci. Adv.* **7**, eabk3283 (2021).
41. Sun, L., Qian, J. & Chen, X. Ldp-fl: Practical private aggregation in federated learning with local differential privacy. In *IJCAI*, p. 1571–1578 (IJCAI, 2021).
42. Banitalebi-Dehkordi, A. Knowledge distillation for low-power object detection: A simple technique and its extensions for training compact models using unlabeled data. In *ICCV*, p. 769–778 (IEEE, 2021).
43. Warner, S. L. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **60**, 63–69 (1965).
44. Zou, Y., Yu, Z., Liu, X., Kumar, B. & Wang, J. Confidence regularized self-training. In *ICCV*, p. 5982–5991 (IEEE, 2019).
45. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 2278–2324 (IEEE, 1998).
46. Hsu, T.-M. H., Qi, H. & Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv. preprint arXiv:1909.06335* (2019).
47. Vaswani, A. et al. Attention is all you need. In *NeurIPS*, p. 5998–6008 (Curran Associates Inc., 2017).
48. Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis* (University of Tront, 2009).
49. Nayak, G. K., Mopuri, K. R., Shaj, V., Radhakrishnan, V. B. & Chakraborty, A. Zero-shot knowledge distillation in deep networks. In *ICML*, p. 4743–4751 (PMLR, 2019).
50. Zhu, Z., Hong, J. & Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. In *ICML*, p. 12878–12889 (PMLR, 2021).
51. Chawla, A., Yin, H., Molchanov, P. & Alvarez, J. Data-free knowledge distillation for object detection. In *WACV*, p. 3289–3298 (IEEE, 2021).
52. Reisizadeh, A., Farnia, F., Pedarsani, R. & Jadbabaie, A. Robust federated learning: The case of affine distribution shifts. In *NeurIPS*, p. 21554–21565 (Curran Associates Inc., 2020).
53. Xie, C., Chen, M., Chen, P.-Y. & Li, B. Crfl: Certifiably robust federated learning against backdoor attacks. In *ICML*, p. 11372–11382 (PMLR, 2021).
54. Sun, J. et al. Fl-wbc: enhancing robustness against model poisoning attacks in federated learning from a client perspective. In *NeurIPS*, p. 12613–12624 (Curran Associates Inc., 2021).
55. Zhang, X., Gu, H., Fan, L., Chen, K. & Yang, Q. No free lunch theorem for security and utility in federated learning. *ACM Trans. Intell. Syst. Technol.* **14**, 1–35 (2022).
56. Ren, X. et al. High-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security.* p. 2151–2166 (IEEE, 2018).
57. Girgis, A., Data, D., Diggavi, S., Kairouz, P. & Suresh, A. T. Shuffled model of differential privacy in federated learning. In *AIStat*, p. 2521–2529 (PMLR, 2021).
58. Qi, T. et al. Differentially private knowledge transfer for federated learning. *PrivateKT* https://doi.org/10.5281/zenodo.7904770 (2023).

Acknowledgements

This work was supported by the National Key Research and Development Program of China 2022YFC3302104 (Y.H.), the National Natural Science Foundation of China under Grant numbers 82090053 (Y.H.), Tsinghua University Initiative Scientific Research Program of Precision Medicine under Grant number 2022ZLA0073 (Y.H.), and Tsinghua-Toyota Joint Research Fund 20213930033 (Y.H.).

Author contributions

Y.H. coordinated the research project and supervised the project with assistance from X.X. T.Q., C.W. and F.W. conceived the idea of this work. T.Q. and C.W. implemented the models for experiments. T.Q., C.W. and F.W. analysed the results. T.Q. and L.H. conducted experiments for the revised paper. All authors contributed to the writing and revision of this manuscript. C.W. was involved in this work while pursuing a Ph.D. at Tsinghua University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-38794-x.

Correspondence and requests for materials should be addressed to Fangzhao Wu, Chuhan Wu or Yongfeng Huang.

Peer review information *Nature Communications* thanks the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.