

Aidmar Wainakh\*, Fabrizio Ventola, Till Müßig, Jens Keim, Carlos Garcia Cordero, Ephraim Zimmer\*, Tim Grube, Kristian Kersting, and Max Mühlhäuser

# User-Level Label Leakage from Gradients in Federated Learning

**Abstract:** Federated learning enables multiple users to build a joint model by sharing their model updates (gradients), while their raw data remains local on their devices. In contrast to the common belief that this provides privacy benefits, we here add to the very recent results on privacy risks when sharing gradients. Specifically, we investigate Label Leakage from Gradients (LLG), a novel attack to extract the labels of the users’ training data from their shared gradients. The attack exploits the direction and magnitude of gradients to determine the presence or absence of any label. LLG is simple yet effective, capable of leaking potential sensitive information represented by labels, and scales well to arbitrary batch sizes and multiple classes. We mathematically and empirically demonstrate the validity of the attack under different settings. Moreover, empirical results show that LLG successfully extracts labels with high accuracy at the early stages of model training. We also discuss different defense mechanisms against such leakage. Our findings suggest that gradient compression is a practical technique to mitigate the attack.

**Keywords:** Label leakage, federated learning, gradient attack, privacy attack.

DOI 10.2478/popets-2022-0043

Received 2021-08-31; revised 2021-12-15; accepted 2021-12-16.

---

**\*Corresponding Author: Aidmar Wainakh:** Telecooperation Lab, Technical University of Darmstadt, E-mail: wainakh@tk.tu-darmstadt.de

**Fabrizio Ventola:** Artificial Intelligence and Machine Learning Lab, Technical University of Darmstadt, E-mail: ventola@cs.tu-darmstadt.de

**Till Müßig:** Technical University of Darmstadt, E-mail: till.muessig@stud.tu-darmstadt.de

**Jens Keim:** Technical University of Darmstadt, E-mail: jens.keim@stud.tu-darmstadt.de

**Carlos Garcia Cordero:** Telecooperation Lab, Technical University of Darmstadt, E-mail: garcia@tk.tu-darmstadt.de

**\*Corresponding Author: Ephraim Zimmer:** Telecooperation Lab, Technical University of Darmstadt, E-mail: zimmer@privacy-trust.tu-darmstadt.de

**Tim Grube:** Telecooperation Lab, Technical University of Darmstadt, E-mail: grube@tk.tu-darmstadt.de

## 1 Introduction

In an increasingly interconnected world, the abundance of data and user information has brought Machine Learning (ML) techniques into daily life and many services. Arguably, the most common ML approaches work in a centralized fashion, typically requiring large amounts of user data to be collected and processed by central service providers. This data can be of sensitive nature, raising concerns about the handling of data in accordance with user expectations and privacy regulations (e.g., the European General Data Protection Regulation, GDPR).

Federated Learning (FL) is an emerging ML setting that allegedly enables service providers *and* users to utilize the power of ML without exposing the user’s personal information. The general principle of FL consists of cooperating to train an ML model in a distributed way. Users are given a model, which they can locally train with their sensitive data. Afterwards, users only share the model gradients of their training endeavors with a central server. The users’ gradients are aggregated to establish the joint model [27]. This general principle is currently believed to reduce the impact on users’ privacy compared to the classical centralized ML setting, since personal information does not leave the user, and sharing learning gradients does not supposedly reveal information about the user [41].

However, a considerable number of recent works have shown that gradients can be exploited to reconstruct the users’ training data in FL [3, 12, 35, 40]; while protecting the users’ ground-truth labels from possible leakage has received only limited attention [21, 38, 41], mainly focusing on gradients generated from a small number of data samples (small batches) or binary classification tasks. Label leakage, however, is a considerable

---

**Kristian Kersting:** Artificial Intelligence and Machine Learning Lab, Technical University of Darmstadt, E-mail: kersting@cs.tu-darmstadt.de

**Max Mühlhäuser:** Telecooperation Lab, Technical University of Darmstadt, E-mail: max@tk.tu-darmstadt.de

risk for FL. Both, FL as well as the more superordinate setting of distributed ML are used in many applications where labels can contain highly sensitive information. For example, in the medical sector, hospitals employ distributed learning to collaboratively build ML models for disease diagnosis and prediction [11, 17]. In some cases, the medical data is collected directly from the patients’ personal devices [8], e.g., mobile phones [9], where an application of FL could introduce many potential benefits. Building models in this and many other settings, while maintaining the users’ privacy, would be crucial. Leaking the labels of the users’ data might disclose their diseases, which is a severe violation of privacy. It is essential to highlight this issue and explore to what extent gradients can leak information about labels. For this purpose, developing privacy attacks that exploit gradients is of high importance in order to foster research and development on the mitigation of respective privacy risks.

Triggered by this, we investigate Label Leakage from Gradients (LLG), a novel attack to extract ground-truth labels from shared gradients trained with mini-batch stochastic gradient descent (SGD) for multi-class classification. LLG is based on a combination of mathematical proofs and heuristics derived empirically. The attack exploits two properties that the gradients of the last layer of a neural network have: **(P1)** The direction of these gradients indicates whether a label is part of the training batch. **(P2)** The gradient magnitude can hint towards the number of occurrences of a label in the batch. Here, we formalize these properties, provide their mathematical proofs, study an extended threat model, and conduct an extensive evaluation, as follows.

- We consider four benchmark datasets, namely, MNIST, SVHN, CIFAR-100, and CelebA . Results show that LLG achieves high success rate despite the datasets having different classification targets and complexity levels.
- We consider two FL algorithms, namely, FedSGD and FedAvg [27]. Results show that for untrained models LLG is more effective under FedSGD, yet poses a serious threat to expose labels under FedAvg as well.
- We study LLG considering different capabilities of the adversary. Experiments demonstrate that an adversary with an auxiliary dataset, which is similar to the training dataset, can adequately extract labels with an accuracy of  $> 98\%$  at the early stage of the model training under the FedSGD algorithm .
- We show that the simple LLG attack can outperform one of the state-of-the-art optimization-based attacks, Deep Leakage from Gradients (DLG) [41], under several settings. Furthermore, LLG is orders of magnitude faster than DLG.
- We also investigate the effectiveness of the attack on various model architectures including simple Convolutional Neural Network (CNN), LeNet [20], and ResNet20 [16]. Results suggest that LLG is not highly sensitive to the complexity of the model architecture.
- We illustrate the influence of the model convergence status on LLG. Findings reveal that LLG can perform best at the early stages of training and still demonstrates information leakage in well-trained models.
- Finally, we test LLG against two defense mechanisms: noisy gradients and gradient compression (pruning). Results show that gradient compression with  $\geq 80\%$  compression ratio can render the attack ineffective.

In this work, we focus on the FL and distributed ML settings because the surface of the attacks against gradients is much wider compared with the centralized training approach. However, LLG can be applied in other scenarios where the gradients of a target user are accessible by an adversary.

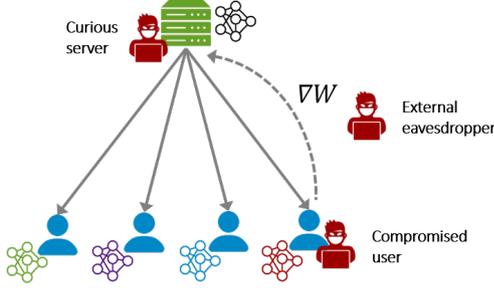
We proceed as follows. We start off by reviewing the background and our problem setting in Section 2. Next, in Section 3, we present related work on information leakage from gradients. We elaborate on our findings regarding gradients properties in Section 4. The attack is then explained in Section 5. Before concluding, we present the results of our evaluation in Section 6.

## 2 Background

In this section, we present the fundamentals of neural networks and FL. Then, we describe our problem setting and the threat model.

### 2.1 Neural networks

Neural Networks (NNs) are a subset of ML algorithms; an NN is comprised of layers of nodes (neurons) including an input layer, one or more hidden layers, and an output layer. The neurons are connected by links associ-



**Fig. 1.** Federated learning overview with three potential adversary access points (in red). Gradients are generated by individual users and shared with a central server. An adversary with access to these gradients can exploit them to estimate the presence and frequency of labels, which can be, e.g., a result of a medical imaging technique for disease prediction.

ated with weights  $\mathbf{W}$ . The NN model can be used for a variety of tasks, e.g., regression analysis, classification, and clustering. In the case of classification, for example, the task of the model  $\hat{f}$  is to approximate the function  $f(\mathbf{x}) = y$  where  $y$  is the class label of a multidimensional data sample  $\mathbf{x}$ , e.g., an image—matrix of pixels. To fulfill this task, the model is trained by optimizing the weights  $\mathbf{W}$  using a loss function  $l$  and training data consisting of input data  $\mathbf{x}_i : i \in [1, N]$  and corresponding labels  $y_i$  in order to solve [12]

$$\min_{\mathbf{W}} \sum_{i=1}^N l_{\mathbf{W}}(\mathbf{x}_i, y_i). \quad (1)$$

Minimizing the loss function can be achieved by applying one of the optimization algorithms. Gradient descent is one of the basic optimization algorithms for finding a local minimum of a differentiable function. This algorithm is based on gradients  $\nabla \mathbf{W}$ , which are the derivative of the loss function w.r.t. the model weights  $\mathbf{W}$ . The core idea is to update the weights through repeated steps  $t$  in the opposite direction of the gradient because this is the direction of steepest descent.

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \eta \nabla \mathbf{W}, \quad (2)$$

where  $\eta$  is the learning rate, which defines the step size for the model updates in the parameter space. An extension of gradient descent, called Minibatch Stochastic Gradient Descent is widely used for training NNs. This algorithm takes a batch of data samples from the training dataset to compute gradients  $\nabla \mathbf{W}$  and, subsequently, updates the weights. The batch size  $B$  is the number of data samples given to the network for each weight update.

## 2.2 Federated learning

Federated Learning (FL) is a machine learning setting that enables a set of  $U$  users to collaboratively train a joint model under the coordination of a central server [18]. For each round  $t$  of the global training process, a subset of users  $K_t \ll U$  is selected to train the model locally on their data. In particular, they optimize the model weights  $\mathbf{W}$  based on the gradients  $\nabla \mathbf{W}$ . Users can take one step of gradient descent (FedSGD [27]) or multiple steps (FedAvg [27]) before sharing the gradients  $\nabla \mathbf{W}$  with the server. The server calculates a weighted average to aggregate the gradients from the  $K_t$  users, and updates the global model

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \eta \sum_{k=1}^{K_t} \frac{v_k}{v} \nabla \mathbf{W}^k, \quad (3)$$

where  $v_k$  is the number of data samples of user  $k$ , and  $v$  is total number of data samples. This process is repeated until the model potentially converges [27]. This setting mitigates a number of privacy risks that are typically associated with conventional machine learning, where all training data should be collected, then used to train a model [18].

## 2.3 Problem setting

We consider a federated setting where  $U$  users jointly train an NN model for a supervised task using either the FedSGD or FedAvg algorithm [27]. For FedSGD, the users train the model locally for one iteration on a batch of their data samples and labels. In FedAvg, each user trains the model for several iterations (multiple batches). We assume the users to be honest, i.e., they train the model with real data and correct labels. Then, the users share the gradients resulted from the local training with the server. We assume that the model consists of  $L$  layers and is trained with cross-entropy loss [14] over one-hot labels for a multi-class classification task. For the studied attack, we focus on the gradients  $\nabla \mathbf{W}_L$  w.r.t. the last-layer weights  $\mathbf{W}_L$  (between the output layer and the layer before), where  $\mathbf{W}_L \in \mathbb{R}^{n \times h}$ :  $n$  is the total number of classes and  $h$  is the number of neurons in layer  $L-1$ . The gradient vector  $\nabla \mathbf{W}_L^i$  represents gradients connected to label  $i$  on the output layer. We note  $g_i$  to refer to the sum of  $\nabla \mathbf{W}_L^i$  elements:  $g_i = \mathbf{1}^T \cdot \nabla \mathbf{W}_L^i$ .

## 2.4 Threat model

We assume that an adversary applies the attack against the shared gradients of one target user. The adversary analyzes the gradients to infer the number of label occurrences in the user’s input data. In FedSGD, this concerns one batch, while in FedAvg, the data consists of multiple batches. Thus, the more information is carried by the gradients on the labels, the higher is the privacy risk. At the same time, the shared gradients need to reflect the training data of the users to optimize the joint model, i.e., to achieve the learning objective. As a result, the learning objective and the depicted privacy risk are mutually related to the information carried by the gradients. Even though it might seem as a paradox, our work is an attempt to focus on and mitigate the privacy risk imposed by gradient sharing without jeopardizing the learning objective of FL and the model accuracy. Next, we define our threat model w.r.t. three aspects: adversary access point, mode, and observation.

**Access point.** The distributed nature of FL increases the attack surface as shown in Figure 1. An adversary might be able to access the gradients by compromising the user’s device as the gradients are calculated on the user side before sharing with the server. We assume that the user’s device can be compromised partially, such that the adversary has no access to the training data or labels [36]. Such scenario can apply, for example, to several online ML applications, where the training data is not stored but used for training on-the-fly. In these cases, compromising a device during or after the training phase would not grant the adversary full access to the training data, while still providing access to the model and possibly the gradients. Other scenarios might exploit a vulnerability in the implementation of the network protocols/interface, such that an adversary accesses only the I/O data. The server also can access the gradients of an individual user, in case no secure aggregation [6] or other protection techniques are used. In addition, if the connection between the server and the users is not secure, the gradients might be intercepted by an external eavesdropper.

**Mode.** We assume the adversary to act in a passive mode. The adversary may analyze the gradients to infer information about the users, but without hindering or deviating from the regular training protocol. This adversary mode is widely common in privacy attacks [26, 30, 37, 41], where the focus is on disclosing information rather than disturbing the system.

**Observation.** The adversary might be capable to observe different amounts of information to launch their attack. We consider three possibilities.

1. Shared gradients: the adversary has access only to the shared gradients. This can apply for an external eavesdropper or an adversary with limited access to the user’s device.
2. White-box model: in addition to the gradients, the adversary is aware of the model architecture and parameters. In the case of a curious server or compromised user, the adversary might have this kind of information.
3. Auxiliary knowledge: the adversary has access to all the aforementioned information and additionally to an auxiliary dataset. This dataset contains data samples of the same classes as the original training dataset. This is a common scenario in real-world cases, given that NNs need a considerable amount of labeled data for training to perform accurately. Labeled data is usually expensive and a typical adopted strategy is to train the model on the publicly available datasets and, eventually, fine-tuning the model on ad-hoc data. Therefore, it is often easy to have access to a big part of the training data.

## 3 Related work

Although the training data is not disclosed to other parties in FL, several works in the literature showed that the data and ground-truth labels can be reconstructed by exploiting the shared gradients. Next, we present existing (1) data reconstruction attacks and (2) label extraction attacks.

### 3.1 Data reconstruction

Aono et al. [2, 3] are the first to discuss reconstructing data from gradients and illustrate its feasibility on a simple NN with a training batch of one sample. The authors closely examined the mathematical definition of the gradients shared with the central server as proposed in [32]. With the help of four examples, they showed how the relationship between the input data, which is unknown to the server, and the gradients can be exploited in order to leak at least some information about the unknown input. Wang et al. [35] moved on to generative attacks, leveraging a Generative Adversarial Network (GAN) to reconstruct the input data in a CNN.

Instead of training the discriminator of the GAN on the server side with real user data, the authors observed, that locally training a shared model on each user like in the FL setting is equivalent. Thus, obtaining the user updates effectively yields updates to the discriminator for each user. The generator of the GAN then is trained on the server side to generate samples indistinguishable from real user samples, which approaches the private training data.

In contrast, Zhu et al. [41] introduced an optimization-based attack; the attacker generates dummy input data and output labels, then optimizes them using L-BFGS [24] to generate dummy gradients that match the shared ones. By that, the dummy data and labels converge to the real data and labels used by the participants in the training process. Instead of using the euclidean distance as a cost function and L-BFGS, Geiping et al. [12] proposed using cosine similarity and the Adam optimization algorithm. They demonstrated that their attack is effective on trained and untrained models, also on deep networks and shallow ones. Furthermore, they proved that the input to any fully-connected layer can be reconstructed regardless of the remaining network architecture. Wei et al. [36] provided a framework for evaluating the reconstruction attacks and discussed the impact of multiple factors (e.g., activation and loss functions, optimizer, batch size) on the cost and effectiveness of these attacks. Qian et al. [31] theoretically analyzed the limits of [41] considering fully-connected NNs and vanilla CNNs. They also proposed a new initialization mechanism to speed up the attack convergence. Unlike previous approaches, Enthoven et al. [10] introduced an analytical attack that exploits fully-connected layers to reconstruct the input data on the server side, and they extend this exploitation to CNNs. Recently, Zhu et al. [40] proposed a recursive closed-form attack. They demonstrated that one can reconstruct data from gradients by recursively solving a sequence of systems of linear equations. Overall, all the aforementioned attacks, except for [41], only focus on reconstructing the input training data while overlooking the leakage of data labels, which can be of high sensitivity. In our research, inspired by the mathematical foundations used in these attacks, we shed more light on the potential vulnerability of label leakage in FL and distributed learning.

### 3.2 Label extraction

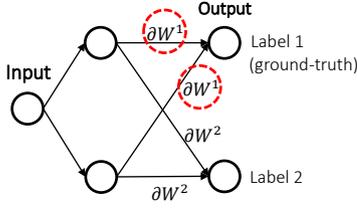
While the data reconstruction attacks attracted considerable attention in the research community, a very limited number of approaches were proposed to extract the ground-truth labels from gradients.

In the work of Zhu et al. [41], the ground-truth labels are extracted as part of their optimization approach. However, the approach requires a learning phase where the model is sensitive to the weight initialization and needs attentive hyperparameter selection. Yet, it can be hard to converge in some cases. Moreover, it was found to extract wrong labels frequently [38] and it is effective only for gradients aggregated from a batch size  $\leq 8$  [28]. Actually, Zhao et al. [38] proposed a more reliable analytical approach to extract the ground-truth labels by exploiting the direction of the gradients. The authors demonstrated that the gradients of classification (cross-entropy loss) w.r.t. the last layer weights have negative values for the correct labels. Thus, detecting the negative gradients is sufficient to extract correct labels. However, their approach is limited to one-sample batch, which is uncommon in real-world applications of FL, where users typically have multiple data samples and train the model on these samples (a bunch of them at least) before sharing the gradients with the server. Wainakh et al. [34], in a short paper (4 pages), introduced a basic idea to extend the attack of [38] for bigger batches, however, their work lacks formalization and thorough evaluation to substantially support the validity of the approach. Li et al. [21] proposed also an analytical approach based on the observation that the gradient norms of a particular class are generally larger than the others. However, their approach is tailored for vertical split learning rather than FL, and it is valid only for a binary classification task.

Overall, the existing approaches are not well generalized to arbitrary batch sizes nor number of classes. Moreover, the influence of different model architectures on these approaches is yet to be investigated. In our work, we take into account these issues by evaluating the LLG attack on a variety of batch sizes and datasets with various numbers of classes, and we involve several model architectures.

## 4 Gradient analysis

In gradient descent optimization, the values of the gradient determine how the parameters of a model need to



**Fig. 2.** Graphical representation of a basic NN model and the gradients  $\nabla \mathbf{W}_L^i$  of the last layer  $L$ . For simplicity, the input layer is represented by a single neuron.

be adjusted to minimize the loss function. Through an empirical analysis, we carefully derive two properties for the sign and magnitude of the gradients that indicate the ground-truth labels. In this section, we formalize these properties, next, in Section 5, we use them as a base to launch the attack.

**Property 1.** For label  $i$  and last layer  $L$  in an NN model with a non-negative activation function, when  $\nabla \mathbf{W}_L^i < 0$ , label  $i$  is present in the training batch on which gradient descent was applied<sup>1</sup>.

*Proof.* We consider an NN model for a classification task. The model is trained using the cross-entropy loss over labels encoded with a one-hot encoding. This loss function  $l$  is defined as

$$l(\mathbf{x}, c) = -\ln \frac{e^{y_c}}{\sum_j e^{y_j}}, \quad (4)$$

where  $\mathbf{x}$  is a multidimensional input instance and  $c$  represents the ground-truth label of  $\mathbf{x}$ . While,  $\mathbf{y} = [y_1, y_2, \dots, y_n]$  is the output vector of the model where each  $y_i \in \mathbf{y}$  is the score predicted for the  $i^{\text{th}}$  class,  $y_c$  is the score assigned to the ground-truth label, and  $n$  is the total number of classes. A graphical representation of a simple NN model and its gradients of the last layer is depicted in Figure 2.

Given a batch size  $B$ , we have a set  $\mathbf{X}$  of  $B$  samples and the set of their labels  $C$ . Thus, we can define a training batch as a set composed of the pairs  $\{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_B, c_B)\}$ . Therefore, we can redefine the loss function as the loss  $l(\mathbf{x}, c)$  of Eq. (4) averaged over a batch of  $B$  labeled samples

$$l(\mathbf{X}, C) = -\frac{1}{B} \sum_{k=1}^B \ln \frac{e^{y_{c(k)}}}{\sum_j e^{y_{j(k)}}}, \quad (5)$$

where  $c(k)$  is the ground-truth label for the  $k^{\text{th}}$  sample in the batch, and  $y_{c(k)}$  is the corresponding output score when  $\mathbf{x}_k$  is given as input to the model. We note that the gradient  $d_i$  of the loss w.r.t. an output  $y_i$  is

$$d_i = \frac{\partial l(\mathbf{X}, C)}{\partial y_i} = -\frac{1}{B} \sum_{k=1}^B \left( \frac{\partial \ln e^{y_{c(k)}}}{\partial y_i} - \frac{\partial \ln \sum_j e^{y_{j(k)}}}{\partial y_i} \right) \quad (6)$$

$$= -\frac{1}{B} \sum_{k=1}^B \left( \mathbf{1}(i = c(k)) - \frac{e^{y_{i(k)}}}{\sum_j e^{y_{j(k)}}} \right), \quad (7)$$

where  $\mathbf{1}(\alpha = \beta) = 1$  if  $\alpha = \beta$ ,  $\mathbf{1}(\alpha = \beta) = 0$  otherwise.

$$d_i = -\frac{1}{B} \sum_{k=1}^B \mathbf{1}(i = c(k)) + \frac{1}{B} \sum_{k=1}^B \frac{e^{y_{i(k)}}}{\sum_j e^{y_{j(k)}}} \quad (8)$$

$$= -\frac{\lambda_i}{B} + \frac{1}{B} \sum_{k=1}^B \frac{e^{y_{i(k)}}}{\sum_j e^{y_{j(k)}}}, \quad (9)$$

where  $\lambda_i$  is the number of occurrences (frequency) of samples with label  $i$  in the training batch. When  $i \notin C$ ,  $\lambda_i = 0$ , and  $e^{y_i} / \sum_j e^{y_j} \in (0, 1)$ , thus,  $d_i \in (0, 1)$ . Instead, when  $i \in C$ , we have  $-\frac{\lambda_i}{B} \leq d_i \leq 1 - \frac{\lambda_i}{B}$ . Hence, if the gradient  $d_i$  is negative, we can conclude that label  $i \in C$ . Of course, the  $d_i$  value moves in this range accordingly to the status of the network weights optimization, e.g. if  $i \in C$  and the network performs poorly, then,  $d_i$  will be closer to  $-\frac{\lambda_i}{B}$ . However, the gradients  $\mathbf{d}$  w.r.t. the outputs  $\mathbf{y}$  are usually not calculated or shared in FL, but only  $\nabla \mathbf{W}$ , the gradients w.r.t. the model weights  $\mathbf{W}$ . We write the gradient vector  $\nabla \mathbf{W}_L^i$  w.r.t. the weights  $\mathbf{W}_L^i$  connected to the  $i^{\text{th}}$  output representing the  $i^{\text{th}}$  class confidence in the output layer as follows

$$\nabla \mathbf{W}_L^i = \frac{\partial l(\mathbf{X}, C)}{\partial \mathbf{W}_L^i} = \frac{\partial l(\mathbf{X}, C)}{\partial y_i} \cdot \frac{\partial y_i}{\partial \mathbf{W}_L^i} \quad (10)$$

$$= d_i \cdot \frac{\partial (\mathbf{W}_L^i \mathbf{a}_{L-1} + b_L^i)}{\partial \mathbf{W}_L^i} \quad (11)$$

$$= d_i \cdot \mathbf{a}_{L-1}, \quad (12)$$

where  $\mathbf{y} = \mathbf{a}_L$  is the activation function of the output layer,  $b_L^i$  is the bias, and  $y_i = \mathbf{W}_L^i \mathbf{a}_{L-1} + b_L^i$ . When non-negative activation functions (e.g. Sigmoid or ReLU) are used,  $\mathbf{a}_{L-1}$  is non-negative. Consequently,  $\nabla \mathbf{W}_L^i$  and  $d_i$  have the same sign. Considering Eq. (9), we conclude that negative  $\nabla \mathbf{W}_L^i$  indicates that the label  $i$  is present in the ground-truth labels set  $C$  of the training batch. However, a present label can have a positive gradient according to the value of  $d_i$  as discussed earlier.  $\square$

<sup>1</sup> This property is a generalization of the main observation in [38] to batches with arbitrary sizes.

**Property 2.** *In untrained models, the magnitude of the gradient  $g_i = \mathbf{1}^T \cdot \nabla \mathbf{W}_L^i$  is approximately proportional to the number of occurrences  $\lambda_i$  of label  $i$  in the training batch.*

*Proof.* Based on Eq. (12), we have

$$g_i = \mathbf{1}^T \cdot \nabla \mathbf{W}_L^i = d_i (\mathbf{1}^T \cdot \mathbf{a}_{L-1}). \quad (13)$$

We substitute  $d_i$  with its expression from Eq. (9) as follows

$$g_i = \left( -\frac{\lambda_i}{B} + \frac{1}{B} \sum_{k=1}^B \frac{e^{y_i(k)}}{\sum_j e^{y_j(k)}} \right) (\mathbf{1}^T \cdot \mathbf{a}_{L-1}). \quad (14)$$

When  $\sum_{k=1}^B \frac{e^{y_i(k)}}{\sum_j e^{y_j(k)}}$  is close to zero, we can write

$$g_i \approx -\frac{\lambda_i}{B} (\mathbf{1}^T \cdot \mathbf{a}_{L-1}), \quad (15)$$

thus,  $g_i$  is proportional to  $\lambda_i$ . We denote  $m$  to be

$$m = -\frac{\mathbf{1}^T \cdot \mathbf{a}_{L-1}}{B}, \quad (16)$$

therefore,  $g_i \approx \lambda_i m$ . We define the parameter *impact*  $m$  as *the change of the gradient value caused by a single occurrence of a label in the training batch*. This value is negative and constant across labels, thus, label-agnostic.

However, for an untrained model, the value of  $\sum_{k=1}^B \frac{e^{y_i(k)}}{\sum_j e^{y_j(k)}}$  strongly depends on the model weight initialization. The prediction score  $y_i$  can be randomly distributed around an uniform random guess  $P = 1/n$ , which the more classes exist in the dataset, the lower is its value, thus, the aforementioned summation goes closer to zero. In some cases,  $y_i$  might be notably high, although the label  $i$  is not present in the training batch. This comes as a result of misclassification and leads to a positive shift in the gradient values. We call this shift *offset*  $s$ , and based on Eq. (14), we can write

$$s_i = \left( \frac{1}{B} \sum_{k=1}^B \frac{e^{y_i(k)}}{\sum_j e^{y_j(k)}} \right) (\mathbf{1}^T \cdot \mathbf{a}_{L-1}). \quad (17)$$

This offset value varies from a label to another, so it is a label-specific value. Using our defined parameters impact  $m$  and offset  $s_i$ , we can reformulate Eq. (14) as follows  $g_i = \lambda_i m + s_i$ . From this equation, it follows easily that the number of occurrences  $\lambda_i$  of label  $i$  can be derived from the parameters  $m$ ,  $s_i$ , and  $g_i$ .  $\square$

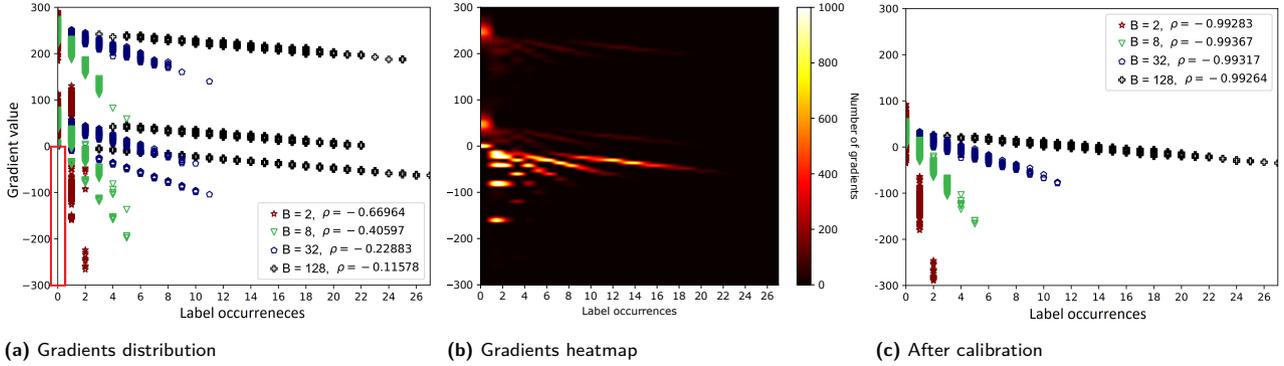
To demonstrate the two gradient properties, we randomly initialize the weights of a CNN composed by

three convolutional layers. Then, we check the gradients  $g_i$  by evaluating the network on a batch of samples taken from the MNIST dataset [20], which contains 10 classes. We repeat the experiment 1,000 times with different batch sizes  $B \in \{2, 8, 32, 128\}$ . Figure 3 (a) depicts the distribution of the resulting gradients, where each data point represents the gradient value of one label in one experiment. The y-axis shows the gradient values and the x-axis represents the number of occurrences for the corresponding label  $i : \forall i \in [1, n]$ .

We can see that there are no negative gradients at  $\lambda_i = 0$  (framed in red), i.e., the negative gradients always correspond to an existing label in the batch  $\lambda_i > 0$ , which confirms Property 1. For all the batch sizes, we notice that the values of the gradients decrease consistently with the increase of the occurrences. This, in turn, confirms Property 2 and our definition of the impact parameter. We also observe that the decrease of gradient values is roughly constant regardless of the label. This confirms the impact being label-agnostic, as we described earlier. Furthermore, we notice that the magnitude of the impact is negatively correlated with the batch size. Meaning, the more samples are present in a batch, the smaller are the changes of the gradients for a different number of occurrences. This is also clear from the definition of impact in Eq. (16). We also can see that there are positive gradients that correspond to  $\lambda_i > 0$ . The positive value of these gradients is mainly caused by the offset  $s_i$  defined in Eq. (17). To illustrate their ratio, we depict a heatmap in Figure 3 (b). We observe that only part of the gradients (18%) are positive, i.e., shifted by the offset, while the majority of the gradients have negative values when the corresponding labels are present in the batch. In Section 5.1, we describe our methods to estimate the offset and elaborate on Figure 3 (c).

## 5 Label extraction

In this section, we present Label Leakage from Gradients (LLG), to extract the ground-truth labels from shared gradients. We first introduce different methods to estimate the attack parameters, impact and offset. Then, we explain the attack.



**Fig. 3.** Distribution of gradients obtained from a randomly initialized CNN on a batch of samples of MNIST varying the batch size in  $\{2, 8, 32, 128\}$ : (a) the distribution shows the correlation between the gradients and the label occurrences, (b) heatmap shows that the majority of the gradients have negative values when the corresponding label is present in the batch, (c) gradients after calibration exhibit a more prominent correlation with label occurrences. Given this strong correlation, it is possible to accurately estimate the label occurrences in the training batch basing on the gradient values.

## 5.1 Attack parameters estimation

In the light of different threat models, we empirically developed several heuristic methods to estimate the impact and offset.

**Shared gradients.** In this scenario, the adversary has access only to the shared gradients. As mentioned earlier, the impact refers to the change in the value of the gradients corresponding to one occurrence of a label. Our intuition is that a good estimation for the impact is obtained by averaging the gradients over the number of data samples  $|\mathcal{D}|$  used by a user in a training round. For FedSGD,  $|\mathcal{D}| = B$  the batch size, while for FedAvg,  $|\mathcal{D}| = \gamma \cdot B$ , where  $\gamma$  is the number of local iterations (batches). Based on Property 1, we know that all negative gradients are indeed indicating existing labels in the training samples. Therefore, we average only the gradients with negative values. Consequently, this average is an underestimation since some gradients may be positive because they are shifted with an offset. We empirically observed that multiplying by a factor that depends on the total number of classes  $n$  is a good additive correction, precisely, we multiply by  $(1 + 1/n)$ . Thus, we estimate the impact  $m$  as follows

$$m = \frac{1}{|\mathcal{D}|} \sum_{i: g_i < 0} (g_i) \left(1 + \frac{1}{n}\right). \quad (18)$$

For this threat model, the offset  $s_i$  cannot be estimated, thus, considered to be zero in the attack.

**White-box model.** When the adversary additionally has access to the model architecture and parameters, they can use it to generate more gradients and gain

more insights about the behavior of the gradients in this model. Consequently, better estimations for the impact and offset can be achieved. The approximation in Eq. (15) indicates that the impact  $m$  can be estimated if the gradient  $g_i$  and number of occurrences  $\lambda_i$  are known, regardless the quality of the input data. Thus, dummy data samples, e.g., dummy images of zeros (black), ones (white), or random pixels, can be used to generate  $g_i$  under known  $\lambda_i$ . More precisely, we form a collection of dummy batches, each batch contains data samples assigned to one label  $i$ . For impact estimation, we pass these batches to a shadow model (a copy of the original model), one at a time, and calculate the average  $\bar{g}_i$  for all the batches corresponding to each label  $i \in [1, n]$ . Then, we average over all classes  $n$  and the batch size  $B$  as follows

$$m = \frac{1}{nB} \sum_{i=1}^n (\bar{g}_i) \left(1 + \frac{1}{n}\right). \quad (19)$$

As mentioned earlier, we assume the offset  $s_i$  to be an approximation of misclassification penalties, when the model mistakenly predicts  $i$  to be ground-truth. These penalties are mainly related to the status of the model weights, which can be biased to specific classes. Based on this intuition, we estimate the offset  $s_i$  by passing batches full of other labels  $\forall j \in [1, n]: j \neq i$ , each batch full of one label, one batch per run. We repeat this for various batch sizes, in total of  $z$  runs. In these runs, the gradients of label  $i$  reflect to some extent the misclassification penalties. Therefore, we calculate the mean of these gradients to be our estimated offset, thus, we have

$$s_i = \frac{1}{z} \sum_{k=1}^z (g_{i_k}). \quad (20)$$

**Auxiliary knowledge.** In this scenario, the adversary has access to the shared gradients, model, and auxiliary data that contains the same classes as the training dataset. Here, the adversary can follow the same methods of the white-box scenario, however, using real input data instead of dummy data. This in turn is expected to yield better estimations for the impact and offset. The goodness of the auxiliary data, i.e., the similarity of the content and class distribution to the original dataset, might play a role in the quality of the estimations. This aspect can be investigated in further research.

To demonstrate the quality of our offset estimation, we calibrate the gradients of Figure 3 (a) by subtracting the estimated offset and plot the results in Figure 3 (c). We can see how the gradients become mainly negative and strongly correlated with the label occurrences. To measure the correlation, we use the Pearson correlation coefficient  $-1 \leq \rho \leq 1$  [4], which yields, for all the studied batch sizes, values of  $|\rho| > 0.99$ . The calibration process mitigates the effect of the offset and makes the gradient values more consistent, thus, easier to be used for extracting the labels.

## 5.2 Label leakage from gradients attack

LLG extracts the ground-truth labels from gradients by exploiting Property 1 and 2. The attack consists of three main steps summarized in Algorithm 1.

1. We start with extracting the labels based on the negative values of the gradients (Property 1). Thus, the corresponding label of each negative gradient is added to the list of the extracted labels  $E$ . As Property 1 holds firm in our problem setting, we can guarantee 100% correctness of the extracted labels in this step. As preparation for the next step, every time we add a label to  $E$ , we subtract the impact from the corresponding gradient following Property 2 (Lines 1-5).
2. We calibrate the gradients by subtracting the offset. In case the offset is not estimated, it is considered to be zero. This step increases the correlation between the gradient values and label occurrences, which facilitates better label extraction based on these values (Line 7).
3. After calibration, the minimum gradient value (negative with maximum magnitude) is more likely cor-

---

### Algorithm 1: Label Leakage from Gradients

---

**Data:**  $\mathbf{G} = [g_1, \dots, g_n]$ : vector of gradients,  
 $m$ : impact,  $\mathbf{S} = [s_1, \dots, s_n]$ : vector of offsets,  $\mathbf{D}$ : data samples used to generate  $\mathbf{G}$ .

**Result:**  $E$ : list for extracted labels.

```

1 for  $g_i \in \mathbf{G}$  do
2   if  $g_i < 0$  then
3     append  $i$  to  $E$ ;
4      $g_i \leftarrow g_i - m$ ;
5   end
6 end
7  $\mathbf{G} \leftarrow \mathbf{G} - \mathbf{S}$ ;
8 while  $|E| < |\mathbf{D}|$  do
9   Select  $g_i : g_i = \min(\mathbf{G})$ ;
10  append  $i$  to  $E$ ;
11   $g_i \leftarrow g_i - m$ ;
12 end
```

---

responding to a label occurred in the batch (see Figure 3 (c)). Therefore, we select the minimum and add the corresponding label to the extracted labels. We repeat Step (3) until the size of the extracted labels list  $E$  matches the number of data samples  $D$  used to generate the gradients. Assuming that  $|\mathbf{D}|$  is known or can be guessed by the adversary (Lines 8-11).

Finally, the output of the LLG attack is the list of extracted labels  $E$ , precisely, the labels existing in the batch and how many times they occur.

## 6 Empirical evaluation

We evaluate the effectiveness of LLG with varying settings including: different FL algorithms, threat models, model architectures, and model convergence statuses. We also test the robustness of LLG against two defense mechanisms, namely, noisy and compressed gradients. For the sake of simplicity, we refer to  $g_i = \mathbf{1}^T \cdot \nabla \mathbf{W}_L^i$  as the gradient of label  $i$  in the rest of this section. Next, we describe the experimental setting, then we discuss our results. The source code of the experiments can be found in <https://github.com/tklab-tud/LLG>.

## 6.1 Experimental setup

**Default model.** We use a CNN model with three convolutional layers (see Appendix, Table 2) as our default model for a classification task. The activation function is Sigmoid, and we use SGD as an optimizer with learning rate 0.1 and cross-entropy as loss function. We use batches of varying sizes  $B = 2^k : k \in [0, 7]$ . When applying the attack for FedSGD, we feed the model with one batch, and we use  $\gamma = 10$  batches for FedAvg. The label distribution in a batch can be *balanced* or *unbalanced*. For balanced data, the samples of the batch are selected randomly from the dataset. For unbalanced data, we select 50% of the batch samples from one random label  $i$  and 25% from another label  $j$ . The remaining 25% of the batch is chosen randomly. We initialize the model with random weights and repeat each experiment 100 times, then report the mean values for analysis and discussion.

**Datasets.** We conduct our experiments on four widely used benchmark datasets: MNIST [20] consists of 70,000 grey-scale images for handwritten digits, with 10 classes in total. SVHN [29] has 99,289 color images of house numbers with 10 classes. CIFAR-100 [19] contains 60,000 color images with 100 classes. And CelebA [25] is a facial attributes dataset with 202,599 images. In our experiments, we consider only the hair color attribute with 5 classes.

**Threat model.** We assume the users to train the model on real data and correct labels. The adversary has access to the shared gradients of only one target user. We consider three different scenarios for the observation capabilities of the adversary (see Section 2.3). Based on these scenarios, the estimation of the impact and offset parameters differs (see Section 5.1), while the same attack applies for all. We refer to the application of the attack under these different scenarios as follows:

1. *LLG* for accessing only the shared gradients scenario.
2. *LLG\** for the white-box model, where we employ various dummy images to estimate the impact and offset. Empirically, we observed the dummy images with which the attack achieves better performance on each dataset. This resulted in using zeros (black) images for MNIST, random pixels for SVHN, ones (white) for CIFAR, and zeros (black) for CelebA.
3. *LLG+* for auxiliary knowledge, where it is assumed that the adversary has access to auxiliary data that contains 10 batches of images from each class.

**Metrics.** To measure the attack effectiveness, we use the attack success rate (ASR) metric [36], which is expressed as the ratio of the correctly extracted labels over

the total number of the extracted labels. We also employed the Hellinger distance [7] to measure the distance between the distribution of the extracted labels and the ground-truth. However, during our experiments, we observed that both aforementioned metrics yielded very similar measurements, therefore, we present our results only with the ASR metric.

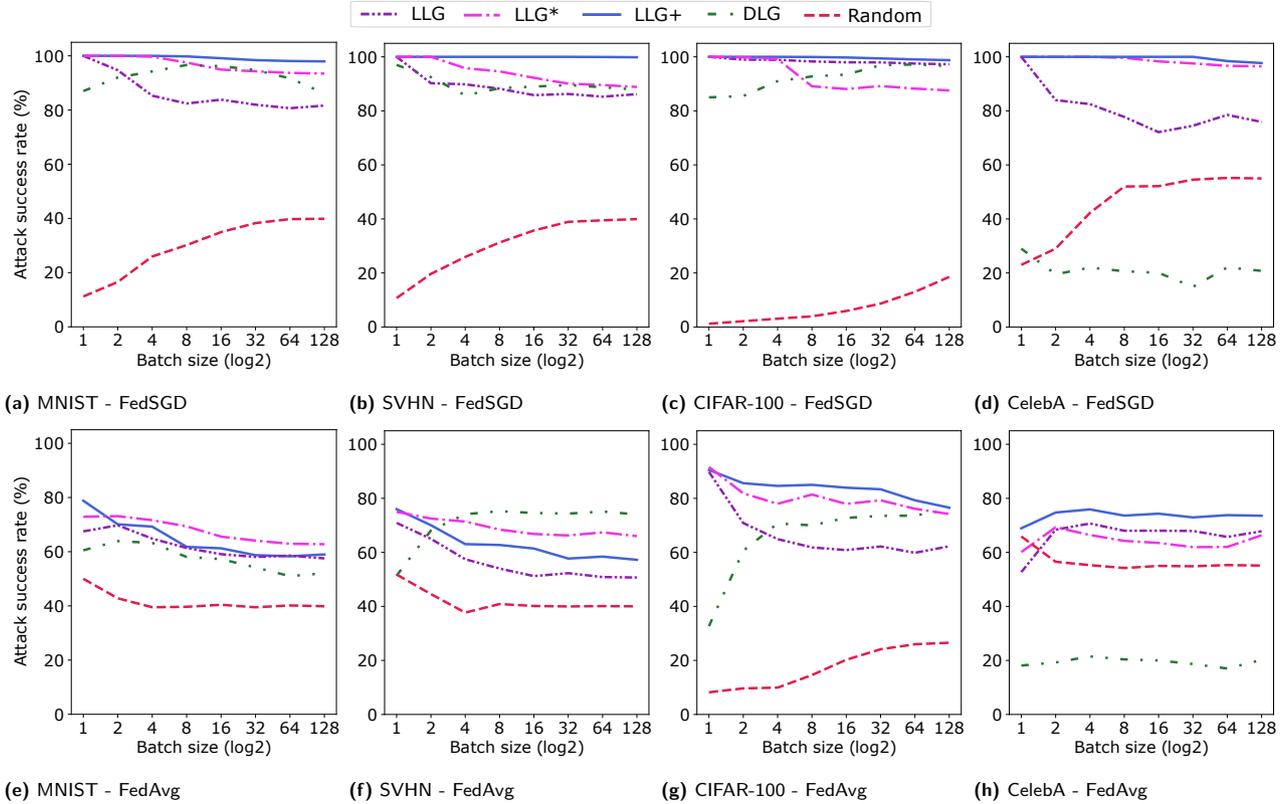
**Baselines.** We compare LLG with two baselines. First, the DLG attack [41], which aims to reconstruct the training data and labels using an optimization approach. For our experiments, we run DLG for 100 iterations and focus only on the label reconstruction results. We used the DLG implementation provided by Zhao et al. [38]<sup>2</sup>. Second, we consider a uniform distribution-based random guess as a baseline. An adversary without any shared gradients might partially succeed in guessing the existing labels frequency by assuming that the labels distribute uniformly, especially in the case of large balanced batches. The random guess serves as a risk assessment curve. Having any attack performing better than the random guess means that there is information leakage.

## 6.2 Attack success rate

We run our experiments under two FL algorithms, namely, FedSGD and FedAvg [27]. For FedSGD, we pass one batch to the model and attack the generated gradients. While for FedAvg, we feed the model with 10 batches and attack the aggregated gradients, i.e., the sum of the gradients over 10 iterations. During our experiments, we observed a very limited difference in the ASR of the LLG attacks for balanced and unbalanced batches. Therefore, and because the unbalanced data is closer to real-world scenarios [27], we focus on presenting the results of the unbalanced data case, while providing a part of the balanced data results (for FedSGD) in the Appendix, Figure 8.

**FedSGD.** Figures 4 (a-d) illustrate the ASR scores (y-axis) with batches of different sizes (x-axis). We can see that all LLG variants show some level of ASR degradation when the batch size increases. However, it appears to be stabilized to some extent for bigger batches, e.g., 64 and 128. This is due to the fact that the first step of the algorithm (see Section 5.2) is based on Prop-

<sup>2</sup> <https://github.com/PatrickZH/Improved-Deep-Leakage-from-Gradients>



**Fig. 4.** Attack success rate of (1) LLG with shared gradients, (2) LLG\* with white-box model, (3) LLG+ with auxiliary knowledge, (4) DLG [41], and (5) random guess on MNIST, SVHN, CIFAR-100, and CelebA. Label extraction is based on gradients generated from passing (1) one batch for FedSGD (first row), and (2) 10 batches for FedAvg (second row), to a randomly initialized CNN. DLG runs for 100 iterations. LLG methods outperform the baselines in most of the cases.

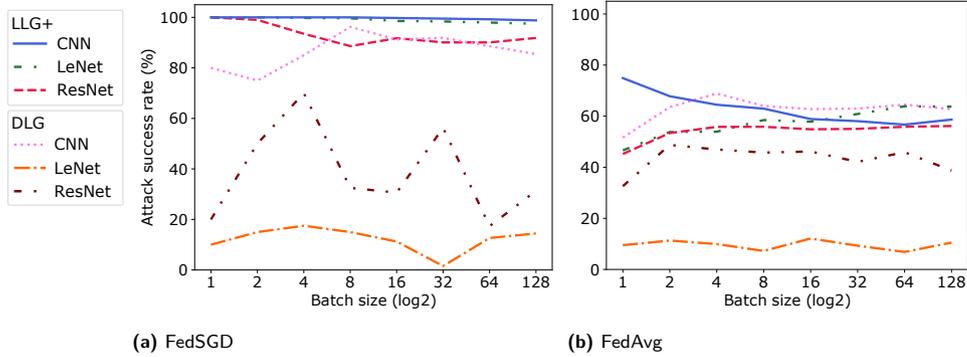
erty 1 and yields 100% correct labels. This step extracts a maximum of  $n$  labels. Thus, its results are dominant when  $B \leq n$ . Where as if  $B > n$ , the second and third steps, which are based on heuristic estimations, contribute more to the final extracted labels. As a result, we notice a degradation of the ASR. However, the different batch sizes do not seem to massively affect the correctness of the results of these steps. This might be explained by the fact that the batch size  $B$  is always considered as a parameter in the heuristic estimations of the impact and offset.

Overall, LLG+ outperforms all the other LLG variants and DLG. The LLG and LLG\* scores range from 100% to a minimum of 77% across the different datasets. Whereas LLG+ remarkably exhibits a high level of stability for various batch sizes and number of classes (in datasets) with an  $ASR > 98\%$ . This mainly reflects the quality of our estimation methods for impact and offset.

In contrast, DLG achieves varying accuracy scores. However, no clear behavior can be concluded w.r.t. the changes in the batch sizes. This might be due to the fact that DLG requires a training phase, which is highly

sensitive to model initialization, i.e., it might fail to converge for some randomly initialized models or might take different amounts of time for reaching a specific accuracy, unlike LLG, which yields more deterministic results, while at the same time being orders of magnitude faster. For example, the execution time of those experiments illustrated in Figure 4 (a) is as follows: LLG 54s, LLG\* 32.2m, LLG+ 14.6m, DLG 17.4h, and Random 50s, using a Tesla GPU V100-SXM3-32GB. It is worth mentioning that LLG\* requires more time than LLG+ due to the dummy images generation.

The ASR of each LLG attack is similar to some extent on MNIST and SVHN respectively. This can be explained by the fact that both datasets have the same number of classes, i.e. 10. On CIFAR-100 (100 classes), interestingly, we notice that LLG performs quite closely to LLG+ (both have  $ASR > 96\%$ ), as shown in Figure 4 (c), while it drops to around 75% on CelebA (5 classes). This observation suggests that LLG performs better for datasets with a bigger number of classes. This can be explained by the fact that LLG solely depends on the quality of the impact parameter, which is derived



**Fig. 5.** Attack success rate of LLG+ and DLG on unbalanced batches of different sizes from MNIST with different model architectures: CNN, LeNet, ResNet20. For FedSGD, LLG+ achieves around 100% accuracy on CNN and LeNet while its accuracy slightly decreases on ResNet20. DLG achieves  $> 80\%$  for most batch sizes on CNN, while drops remarkably on more complex architectures i.e. LeNet and ResNet20. For FedAvg, the ASR of LLG+ is slightly different from architecture to another, while DLG shows higher sensitivity to the architectures.

from Eq. (15) under the assumption that the untrained model performs poorly. This assumption is more valid when the number of classes is bigger, as we explained earlier in the proof of Property 2, Section 4. Therefore, the estimation of the impact yields better results leading to higher ASR.

LLG\*, with its dummy data for the parameter estimation, shows a notable drop on CIFAR-100. It is known that the complexity of CIFAR-100 images is higher than the one of MNIST and SVHN. Therefore, we can conclude that the complexity of the dataset might influence LLG\* in a negative way, while it has no observable effect on LLG and LLG+. For DLG, we notice in Figure 4 (d) a remarkable decrease in accuracy on CelebA. This can be due to the fact that the images are of higher dimensions ( $178 \times 218$ ), unlike the other datasets. Thus, the convergence of the attack is much more difficult.

**FedAvg.** In Figures 4 (e-h), we can see that the ASRs of all the LLG variants considerably decrease comparing with FedSGD, ranging between 55% and 90%. This is expected as the shared gradients are generated from multiple iterations (10 batches). Thus, the correlation between the gradient values and label occurrences is less prominent. In other words, the gradients are accumulated several times over iterations, such that the correlation (Property 1 and 2) become more difficult to detect and exploit. However, the LLG attacks achieve higher ASRs than the random guess on all the datasets, thus, they are still posing a serious threat. The superiority of LLG+ is maintained on CIFAR-100 and CelebA, while in MNIST and SVHN, LLG\* and DLG surprisingly perform the best.

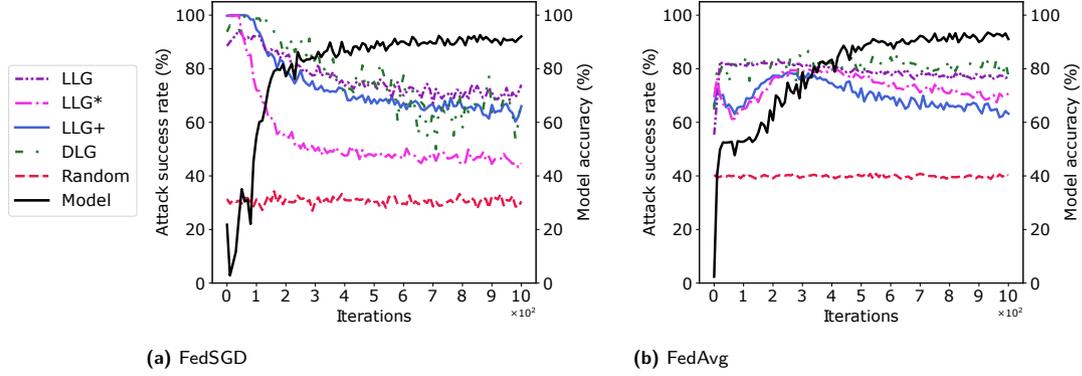
### 6.3 Model architecture

Here, we study the influence of the model architecture on the studied attacks; for that, we consider two models besides our default CNN: (1) LeNet [20], a basic CNN that contains 3 convolutional layers with 2 maximum pooling layers as shown in the Appendix, Table 3. (2) ResNet20 [16], a successful residual architecture with convolutions, which introduces the concept of “identity shortcut connection” that skips one or more layers to avoid the problem of vanishing gradients in deep neural architectures. ResNet20 contains 20 layers in total: 9 convolutional layers, 9 batch normalization, and 2 linear layers.

Both aforementioned architectures, alongside their principal components, namely, convolutions and residual blocks, have achieved and contributed to state-of-the-art results on several classification tasks.

The two main conditions for Property 1 to hold are: (1) using the cross-entropy loss and (2) having a non-negative activation function in the last layer before the output. Thus, we assume the labels extracted in the first step of the attack (see Algorithm 1, Line 1-5) based on this property to be correct regardless of the rest of the model architecture. While the next steps of LLG are based on the impact, offset, and their estimations, which might be of different accuracy from one model architecture to another. To run our analysis, we use MNIST with varying batch sizes and measure the ASR of LLG+ and DLG.

**FedSGD.** As we can see in Figure 5 (a), LLG+ performs best on CNN and LeNet, achieving approximately 100% of success rate, while a degradation starts from



**Fig. 6.** Influence of model convergence status on ASR of LLG, LLG\*, LLG+, DLG, and random guess for CNN with unbalanced batches from MNIST dataset. On the left y-axis it is plotted the attack success rate, while on the right y-axis, it is plotted the model test accuracy. The number of training iterations ( $\times 10^3$ ) is on the x-axis. All different LLG methods achieve remarkable success rates even if the models are well-trained and gradients become less informative.

batches with size  $> 2$  for ResNet20. This is mainly due to the residual blocks in the ResNet20 architecture that prevents the vanishing gradients problem in deep neural networks. In other words, ResNet20 implicitly alters and controls the range of the gradient values in order to not let them vanish (gradients close to zero) or to explode (gradients go towards  $+\infty$ ) during training. This manipulates our definitions of the impact and offset parameters in Eq. (16) and (17), thus, directly affecting the attack performance.

On the other hand, DLG shows much higher sensitivity towards the model architecture. As we can see it achieves  $> 80\%$  for most batch sizes on CNN, while dropping remarkably on more complex models, LeNet and ResNet20. Such a strong influence of the model architecture on DLG is expected, as DLG includes an optimization phase, where optimizing complex models typically requires much more iterations.

**FedAvg.** In Figure 5 (b), we observe that under small batch sizes  $B \leq 16$ , the ASR of LLG+ is higher for CNN. While for bigger batches, LLG+ only slightly differs from one model to another. This supports the finding that the model architecture has limited effect on LLG+. In contrast, DLG shows again higher sensitivity with bigger variance of the ASR over the different architectures.

## 6.4 Model convergence status

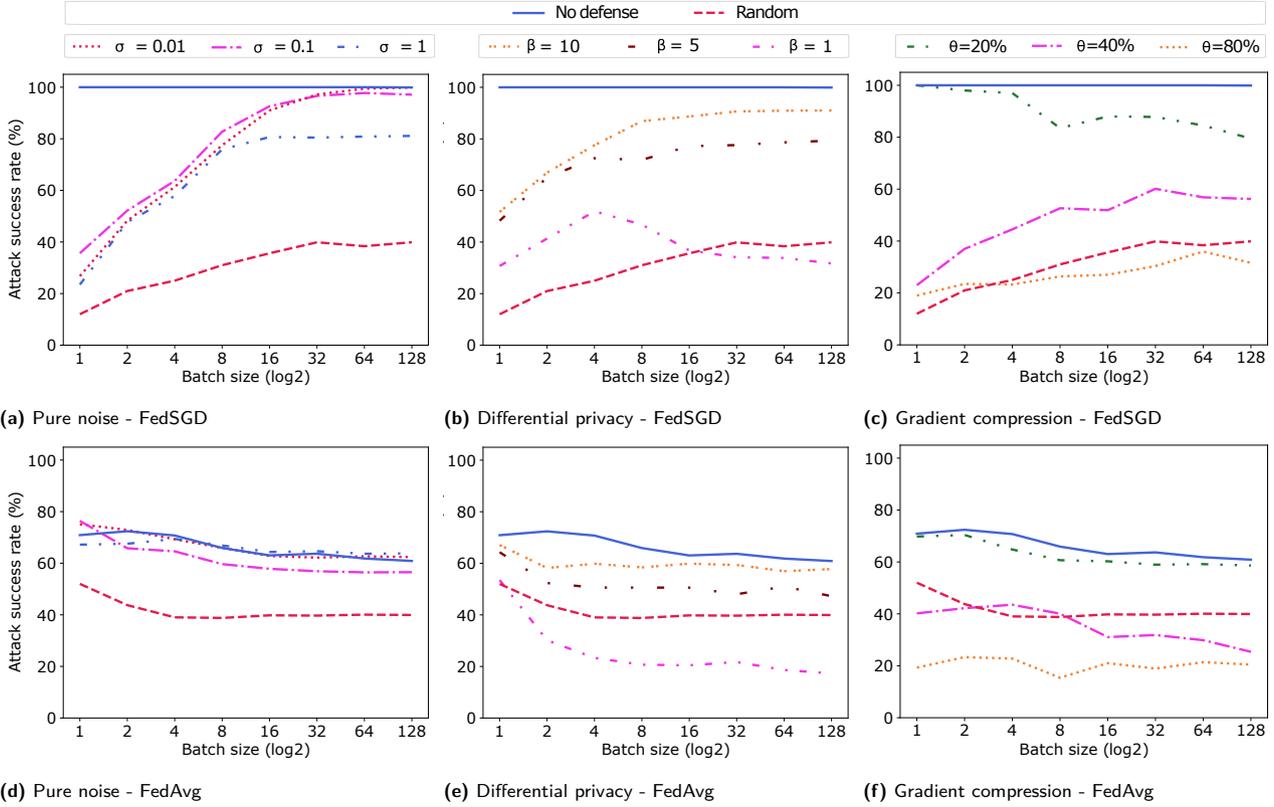
The gradients guide the model towards a local minimum of the loss function. As the model converges to this minimum, the information included in the gradients becomes less prominent. Therefore, we expect the

convergence status of the model to have a strong influence on the attack effectiveness. All the previous experiments are conducted in one communication round, i.e., the gradients are generated and shared with the server only once. In this section, we go further with training the model and observe the implications on the attack.

We train the model in a federated setting, where the data of MNIST is distributed among 750 users, each has 80 unbalanced data samples. The server selects randomly 100 users for every communication round to train the global model locally and share their gradients. The CNN model is trained with batches of size 8 for  $10^3$  iterations. We chose the batch size 8 to be able to apply the DLG attack in its most effective setting  $B \leq 8$  [41]. In every communication round, we attack the shared gradients of one target user (victim) with DLG and LLG variants, where the impact and offset are estimated dynamically. Figures 6 (a,b) depict the attacks ASR (on the left y-axis) versus the model accuracy at testing time (on the right y-axis), while the x-axis represents the number of training iterations.

**FedSGD.** We can see in Figure 6 (a) that the growth of the model accuracy incurs a notable decrease of the ASR for all LLGs. However, although the model converges close to 93% accuracy, LLG and LLG+ keep achieving  $ASR > 60\%$ , considerably higher than the random guess which is around 32%. Meaning, the attacks are still able to take advantage of the reduced information in gradients over the course of the whole training process. Similarly, DLG shows degradation in accuracy, yet it remains effective for well-trained models.

**FedAvg.** Figure 6 (b) shows more stability of ASR over the training process, especially for LLG and DLG.



**Fig. 7.** Effectiveness of different defenses against LLG+ on an unbalanced batch from MNIST with a randomly initialized CNN: (a) defense by adding Gaussian distributed noise to gradients with  $\sigma \in \{0.01, 0.1, 1\}$ , (b) defense by user-side differential privacy with  $\sigma = 0.1$ , and clipping bound  $\beta \in \{1, 5, 10\}$ , (c) defense by pruning gradients with varying compression ratios  $\theta \in \{20\%, 40\%, 80\%\}$ . Pure noise is not successful in eliminating the risk completely, since LLG+ maintains a higher ASR than the random guess. While, differential privacy mitigates the attack with  $\beta = 1$  for FedAvg, and for FedSGD when batch size  $B \geq 16$ . Gradient compression is effective in FedSGD when a high compression ratio ( $\geq 80\%$ ) is used with  $B \geq 4$ . For FedAvg, even the compression ratio 40% with  $B \geq 8$  is an effective defense.

Even in the early stages of the training, the multiple local iterations in FedAvg improve the model accuracy, thus, make the accumulative gradients less informative. Therefore, the attacks start with lower ASR comparing with FedSGD. However, this leads also to mitigating the notable degradation of ASR observed in Figure 6 (a). LLG\* and LLG+ exhibit volatile behavior in the early iterations, where they have an increasing success rate between iteration 100 and 300. Then, they decrease again from 80% to close to 70% and 60%, respectively. Interestingly, DLG maintains a high success rate (around 80%) outperforming the LLG variants in most parts of the training process. This shows that DLG is less sensitive to convergence status under FedAvg and thus can cope with the decreasing amount of information in the gradients. Overall, all the attacks stay effective with  $ASR > 40\%$ , which is the random guess success rate.

## 6.5 Defense mechanisms

As LLG is mainly based on the gradients, thus, sensitive to changes in their values, obfuscating them can be a direct mitigation mechanism. In this section, we use two obfuscation techniques: noisy gradients and gradients compression. We apply these techniques on the user side before sharing the gradients with the server and thus, protecting against external eavesdroppers and curious servers. Then, we attack the gradients of one target user in one communication round for a randomly initialized CNN model (untrained). In general, applying obfuscation techniques incurs a loss in the model accuracy. To cover this aspect, we train the model to convergence under conditions similar to those in Section 6.4 while applying the defenses, and report its accuracy.

### 6.5.1 Noisy gradients

Many researchers consider adding noise to gradients as the de facto standard for privacy-preserving ML [22]. In this experiment, we evaluate LLG+ against two techniques of noise addition: (1) Pure noise: we add noise on gradients before sharing, similar to [36, 41], where no formal privacy properties are guaranteed, and (2) differential privacy: following differentially private FL [13], we clip the gradients to bound their sensitivity, then, we add noise to them. The clipping is defined as  $\nabla \mathbf{W} \leftarrow \nabla \mathbf{W} / \max\left(1, \frac{\|\nabla \mathbf{W}\|_2}{\beta}\right)$ , where  $\beta$  is the gradient norm bound. In both noise addition techniques, we use the Gaussian noise distribution. For pure noise, the standard deviation of the noise distribution is  $\sigma \in \{0.01, 0.1, 1\}$  with central 0. For differential privacy, we use  $\sigma = 0.1$  and varying norm bound  $\beta \in \{1, 5, 10\}$ . We track the privacy loss for the model trained with differential privacy using the moments accountant [1]. For 100 communication rounds and  $\delta = 10^{-5}$ , the privacy budget is estimated  $\epsilon \approx 11.5$ .

**FedSGD.** In Figure 7 (a), we can see that the higher the magnitude of the noise the less accurate the attack. This is expected as the attack partially uses the magnitude of the gradients to infer the labels following Property 2. Interestingly, we observe that the noise has less effect on the attack when the batch size is increasing. We investigated this observation further by inspecting the values of the gradients before and after noise addition. Our empirical analysis showed earlier in Figure 3 (a, b) that the majority of gradients  $g_i$  have values close to zero when they correspond to labels not present in the batch. Adding noise to such small gradient values might lead to flipping their sign, and consequently, disrupting Property 1, which is one of the basis of the attack. For batch sizes  $B < n$  with  $n$  as the number of classes, not all the labels will be present in the batch, so the flipping effect can be prominent on the attack success rate. Whereas, in bigger batches, it is more likely to have more differing labels, thus, their gradients values are not close to zero. As a result, adding a small amount of noise does not lead to sign flipping. This also explains the stability of ASR values when  $B \geq n$ . Overall, adding noise does not eliminate the risk completely while reducing the model accuracy (see Table 1). As we can see, LLG+ maintains higher ASRs than the random guess for all the test noise scales.

Figure 7 (b) shows that adding noise of  $\sigma = 0.1$  with clipping bound  $\beta = 1$ , is an effective defense against

LLG+ for batch sizes  $B > 16$ , where the ASR drops beyond the random guess. However, this leads to a significant drop in the model accuracy as shown in Table 1.

**FedAvg.** Unlike in FedSGD, the magnitude of the pure noise does not have a clear effect on ASR for FedAvg as shown in Figure 7 (d). That is due to the fact that the shared gradients are generated from 10 batches. Thus, the gradient values reflect  $10 \times B$  labels, which is always greater or equal  $n$  for MNIST, where  $n = 10$ . Therefore, it is likely that most of the labels appear in one of the batches at least, consequently, no gradient values will be close to zero. As a result, the pure noise does not impact ASR remarkably, and LLG+ remains effective. In Figure 7 (e), we notice that noise  $\sigma = 0.1$  with bound of  $\beta = 1$  is able to mitigate LLG+, reducing its success rate close to 20% for bigger batch sizes. However, the model accuracy degrades remarkably to 52.5%. Additionally, other differential privacy approaches, e.g., DP-SGD [1] can also be applied and investigated as a defense.

### 6.5.2 Gradient compression

Gradient compression [23, 33] prunes shared gradients with small magnitudes to zero. Pruning some gradients reduces the information that the attack exploits to extract the labels. In this set of experiments, we evaluate LLG+ under various gradient compression ratios  $\theta \in \{20\%, 40\%, 80\%\}$ , i.e.,  $\theta$  denotes the percentage of the gradients to be discarded in each communication round with the server. We use the sparsification approach proposed in [23], where users send only the prominent gradients, i.e., with a magnitude larger than a specific threshold. The threshold is calculated dynamically based on the desired compression ratio. The small gradients are accumulated across multiple communication rounds and sent only when they are large enough.

**FedSGD.** Figure 7 (c) illustrates that when the ratio is  $\leq 20\%$ , there is only a slight effect on the success rate of the attack. When the compression ratio is 80%, we notice that LLG+ becomes completely ineffective for  $B \geq 4$ , dropping below the random guess. Notably, the model accuracy is maintained high 91.9% in this case. Consequently, gradient compression with  $\theta > 80\%$  can practically defend against the attack while producing accurate models.

**FedAvg.** Similar to FedSGD, we observe a limited effect of the ratio  $\theta \leq 20\%$  in Figure 7 (f), whereas the ratio of  $\theta = 40\%$  with  $B \geq 8$  can mitigate the risk of LLG+, as well as  $\theta \geq 80\%$  for any batch size. Under both

	FedSGD (Acc. = 93.3%)			FedAvg (Acc. = 94.5%)		
<b>PN</b> ( $\sigma$ )	0.01	0.1	1	0.01	0.1	1
Acc. (%)	93.4	89.9	$\leq 10.1$	94.6	91.4	$\leq 13.5$
<b>DP</b> ( $\beta$ )	10	5	1	10	5	1
Acc. (%)	89	86.1	$\leq 52.4$	91.2	90.5	52.5
<b>GC</b> ( $\theta\%$ )	20	40	80	20	40	80
Acc. (%)	93.4	93.7	91.9	92.8	91.6	89.3

**Table 1.** Model accuracy while applying the defense mechanisms. **PN**: pure noise, **DP**: differential privacy, **GC**: gradient compression.

compression ratios, the model converges at high accuracy scores, 91.6% and 89.3%, respectively. Additional improvements on the accuracy can be achieved by applying error compensation techniques, such as momentum correction and local gradient clipping, which are proposed in [23].

In addition to the aforementioned defenses, cryptography-based approaches exist [3, 5, 15, 39], which can protect gradients from external eavesdroppers and even curious servers. However, besides the computation and communication overhead introduced by these approaches, they prevent the server from evaluating the utility and benignity of users’ updates.

## 7 Conclusion

We identified and formalized two properties of gradients of the last layer in deep neural network models trained with cross-entropy loss for a classification task. These properties reveal a correlation between gradients and label occurrences in the training batch. We investigate Label Leakage from Gradients (LLG), a novel attack that exploits this correlation and extracts the ground-truth labels from shared gradients in the FedSGD and FedAvg algorithms. We demonstrated the validity of LLG through mathematical proofs and empirical analysis. Results demonstrate the scalability of LLG to arbitrary batch sizes and number of classes. Moreover, we showed the success rate of LLG on various model architectures and in different stages of training. The effectiveness of noisy gradients and gradient compression as defenses was also investigated. Findings suggest gradient compression to be an efficient technique to prevent the attack while maintaining the model accuracy. With this work, we hope to raise the awareness of the privacy risks associated with gradients sharing schemes, encouraging the community and service providers to give careful consideration to security and privacy measures in this con-

text. As future work, we are developing improvements for the attack under FedAvg and against trained models. Additionally, we are investigating the implications of combining LLG with DLG on the overall accuracy of the data reconstruction.

## Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 251805230/GRK2050. *Fabrizio Ventola* and *Kristian Kersting* acknowledge the support by the project “kompAKI - The Competence Center on AI and Labour”, funded by the German Federal Ministry of Education and Research. *Kristian Kersting* acknowledges also the support by “KISTRA – Use of Artificial Intelligence for Early Detection of Crimes”, a project funded by the German Federal Ministry of the Interior, Building and Community, FKZ: 13N15343, and “safeFBDC - Financial Big Data Cluster”, a project funded by the German Federal Ministry for Economics Affairs and Energy as part of the GAIA-x initiative, FKZ: 01MK21002K.

## References

- [1] Martín Abadi, H. Brendan McMahan, Andy Chu, Ilya Mironov, Li Zhang, Ian Goodfellow, and Kunal Talwar. Deep learning with differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security*, 2016.
- [2] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning: Revisited and enhanced. In *International Conference on Applications and Techniques in Information Security*, pages 100–110. Springer, 2017.
- [3] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.
- [4] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.
- [5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.
- [6] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer*

- and Communications Security, pages 1175–1191, 2017.
- [7] Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.
  - [8] Matthew F Daley, Kristin Goddard, Melissa McClung, Arthur Davidson, Gretchen Weiss, Ted Palen, Carsie Nyirenda, Richard Platt, Brooke Courtney, and Marsha E Reichman. Using a handheld device for patient data collection: a pilot for medical countermeasures surveillance. *Public Health Reports*, 131(1):30–34, 2016.
  - [9] Sinead Duane, Meera Tandan, Andrew W Murphy, and Akke Vellinga. Using mobile phones to collect patient data: lessons learned from the simple study. *JMIR research protocols*, 6(4):e61, 2017.
  - [10] David Enthoven and Zaid Al-Ars. Fidel: Reconstructing private training samples from weight updates in federated learning. *arXiv preprint arXiv:2101.00159*, 2021.
  - [11] Mona Flores, Ittai Dayan, Holger Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Abidin, Andrew Liu, Anthony Costa, Bradford Wood, et al. Federated learning used for predicting outcomes in sars-cov-2 patients. *Research Square*, 2021.
  - [12] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
  - [13] R. C. Geyer, T. Klein, and M. Nabi. Differentially Private Federated Learning: A Client Level Perspective. *ArXiv e-prints*, December 2017.
  - [14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
  - [15] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics*, 16(10):6532–6542, 2019.
  - [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
  - [17] Arthur Jochems, Timo M Deist, Issam El Naqa, Marc Kessler, Chuck Mayo, Jackson Reeves, Shruti Jolly, Martha Matuszak, Randall Ten Haken, Johan van Soest, et al. Developing and validating a survival prediction model for nslc patients through distributed learning across 3 countries. *International Journal of Radiation Oncology\* Biology\* Physics*, 99(2):344–352, 2017.
  - [18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.
  - [19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *MIT*, 2009.
  - [20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
  - [21] Oscar Li, Jiankai Sun, Weihao Gao, Hongyi Zhang, Xin Yang, Junyuan Xie, and Chong Wang. Label leakage and protection in two-party split learning. *NeurIPS 2020 Workshop on Scalability, Privacy, and Security in Federated Learning (SpicyFL)*, 2020.
  - [22] Qinbin Li, Zeyi Wen, and Bingsheng He. Federated learning systems: Vision, hype and reality for data privacy and protection. *arXiv preprint arXiv:1907.09693*, 2019.
  - [23] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.
  - [24] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
  - [25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
  - [26] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning. *arXiv preprint arXiv:2010.10152*, 2020.
  - [27] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.
  - [28] Fan Mo, Anastasia Borovykh, Mohammad Malekzadeh, Hamed Haddadi, and Soteris Demetriou. Layer-wise characterization of latent information leakage in federated learning. *arXiv preprint arXiv:2010.08762*, 2020.
  - [29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *CiteSeerX*, 2011.
  - [30] Anastasia Pustozero and Rudolf Mayer. Information leaks in federated learning. In *Proceedings of the Network and Distributed System Security Symposium*, 2020.
  - [31] Jia Qian and Lars Kai Hansen. What can we learn from gradients? *arXiv preprint arXiv:2010.15718*, 2020.
  - [32] Reza Shokri and Vitaly Shmatikov. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15*, pages 1310–1321, New York, NY, USA, 2015. Association for Computing Machinery.
  - [33] Yusuke Tsuzuku, Hiroto Imachi, and Takuya Akiba. Variance-based gradient compression for efficient distributed deep learning. *arXiv preprint arXiv:1802.06058*, 2018.
  - [34] Aidmar Wainakh, Till Müßig, Tim Grube, and Max Mühlhäuser. Label leakage from gradients in distributed machine learning. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–4. IEEE, 2021.
  - [35] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2512–2520. IEEE, 2019.
  - [36] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating client privacy leakages in federated

learning. In *European Symposium on Research in Computer Security*, pages 545–566. Springer, 2020.

- [37] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. Gan enhanced membership inference: A passive local attack in federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2020.
- [38] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.
- [39] Huafei Zhu, Zengxiang Li, Merivyn Cheah, and Rick Siow Mong Goh. Privacy-preserving weighted federated learning within oracle-aided mpc framework. *arXiv preprint arXiv:2003.07630*, 2020.
- [40] Junyi Zhu and Matthew Blaschko. R-gap: Recursive gradient attack on privacy. *arXiv preprint arXiv:2010.07733*, 2020.
- [41] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, pages 14747–14756, 2019.

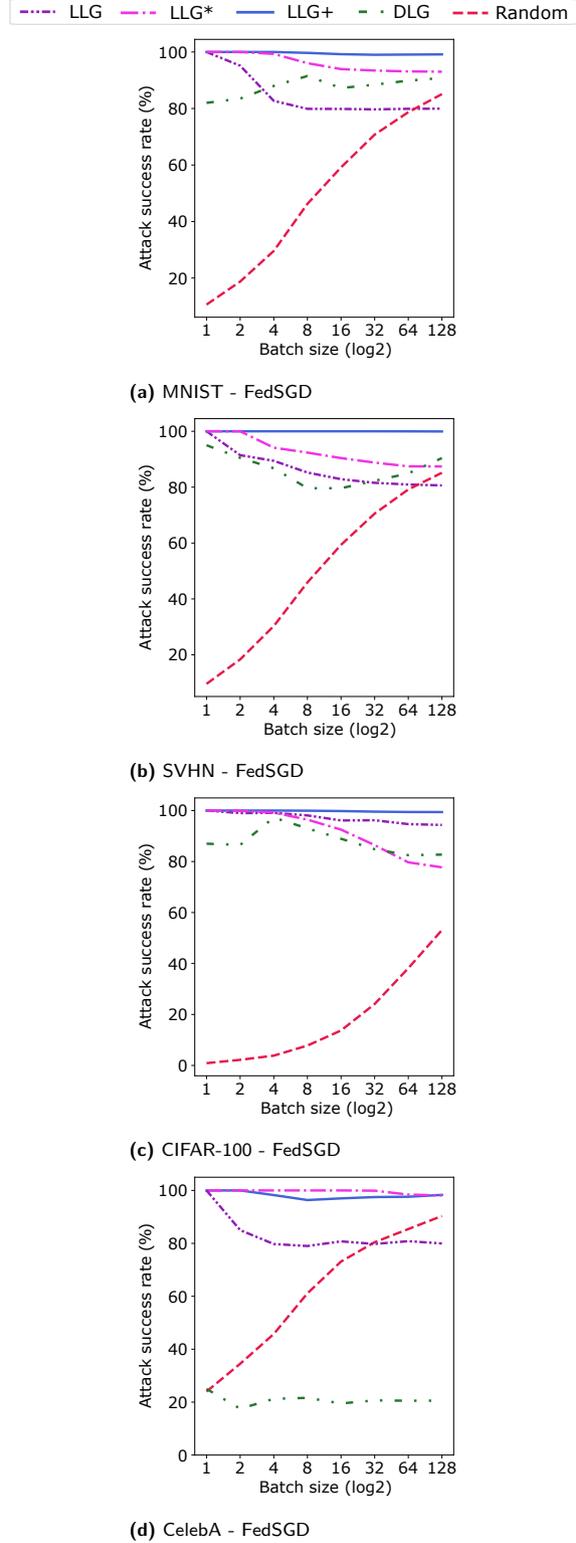
## Appendix

Layer	Size	Activation function
(input)	-	-
Conv 2D	channels $\times$ 12	Sigmoid
Conv 2D	12 $\times$ 12	Sigmoid
Conv 2D	12 $\times$ 12	Sigmoid

**Table 2.** Architecture of CNN, the default model in the experimental setting.

Layer	Size	Activation function
(input)	-	-
Conv 2D	1 $\times$ 6	ReLU
Maxpool	2 $\times$ 2	-
Conv 2D	6 $\times$ 16	ReLU
Maxpool	2	-
Linear	16 $\times$ 6	ReLU
Linear	120 $\times$ 84	ReLU
Linear	84 $\times$ 10	ReLU

**Table 3.** Architecture of LeNet network, a very common architecture adopted in computer vision.



**Fig. 8.** Attack success rate of (1) LLG with shared gradients, (2) LLG\* with white-box model, (3) LLG+ with auxiliary knowledge, (4) DLG [41], and (5) random guess on MNIST, SVHN, CIFAR-100, and CelebA. Label extraction is based on gradients generated from passing one *balanced* batch (FedSGD) to a randomly initialized CNN.