Attacking and Protecting Data Privacy in Edge-Cloud Collaborative Inference Systems

Zecheng He, Tianwei Zhang, Ruby B. Lee, *Fellow, IEEE*

***Abstract*—Benefiting from the advance of Deep Learning technology, IoT devices and systems are becoming more intelligent and multi-functional. They are expected to run various Deep Learning inference tasks with high efficiency and performance. This requirement is challenged by the mismatch between the limited computing capability of edge devices and large-scale Deep Neural Networks. Edge-cloud collaborative systems are then introduced to mitigate this conflict, enabling resource-constrained IoT devices to host arbitrary Deep Learning applications.**

However, the introduction of third-party clouds can bring potential privacy issues to edge computing. In this paper, we conduct a systematic study about the opportunities of attacking and protecting the privacy of edge-cloud collaborative systems. Our contributions are twofold: (1) we first devise a set of new attacks for an untrusted cloud to recover arbitrary inputs fed into the system, even if the attacker has no access to the edge device's data or computations, or permissions to query this system. (2) We empirically demonstrate that solutions that add noise fail to defeat our proposed attacks, and then propose two more effective defense methods. This provides insights and guidelines to develop more privacy-preserving collaborative systems and algorithms.

***Index Terms*—Security and Privacy, Edge-Cloud Computing, Artificial Intelligence, Collaborative Inference**

I. Introduction

Recent years have witnessed the rapid development of Deep Learning (DL) and Internet of Things (IoT) technologies. IoT devices become appealing targets for DL applications. They use various sensors (e.g., cameras, microphones, gyroscopes) to collect data and information from environmental contexts, run the DL applications to interpret sensory data, and make control decisions. The integration of AI and IoT leads to the era of Artificial Intelligence of Things (AIoT), which has significantly changed our daily life: small-scaled AIoT systems are introduced to build smart homes and increase the comfort and quality of life; medium-scale AIoT systems are deployed in warehouses and factories for higher efficiency and automation; large-scale AIoT systems can contribute to the establishment of smart cities.

Deploying deep learning inference applications on commodity edge devices has several challenges. On one hand, an IoT device can collect streaming information at a very high rate (e.g. vehicle detection [2], remote monitoring [3], scene analysis [4] and application trace analysis [5]). This requires the device to run the DL models and analyze the data at a high speed. On the other hand, state-of-the-art DL models are becoming more complicated with larger sizes, making it infeasible for resource-constrained IoT devices to satisfy the performance requirements: the limited computation resources of the device can cause significant latency; the limited storage capacity makes it hard to store a large DNN model; the limited battery capacity causes a critical energy consumption constraint.

To overcome this challenge, one possible approach is to offload the entire DL model and inference computation to the cloud. The edge device sends the input data to the cloud and receives the output. While this can resolve the aforementioned limitations of edge devices, it incurs significant communication costs when sending a large volume of raw data. Besides, there can be privacy breaches of the inference data [6], especially if the input data are highly sensitive like patients' records, and integrity breaches of the model [7], if the cloud is not trusted.

An optimized strategy is to adopt collaborative inference between the edge devices and the cloud [8], [9], [10], [11], [12]. The DL model can be divided into two parts. The first few layers of the network are stored in the local edge device, while the rest are offloaded to a remote cloud. Given an input, the edge device calculates the output of the first layers, sends it to the cloud, and retrieves the final results. This approach can reduce communication costs, as the intermediate output can be designed to be much smaller than the raw input. Such low data transfer bandwidth also achieves lower latency and smaller energy consumption. Collaborative inference makes it feasible and efficient to deploy large-scale intelligent workloads on today's edge platforms.

This paper presents an investigation of inference data privacy in edge-cloud collaborative systems, from the perspectives of attacks and defenses. Prior works all aimed to improve the performance and efficiency of such systems, while ignoring potential security issues. To the best of our knowledge, we are the first to demonstrate the feasibility of input data privacy attacks against cloud-edge collaborative inference systems. The data privacy considered in this paper is the condidentiality of the raw inputs.

Two key questions are considered in this study. The first one is: *if the cloud is malicious or compromised, can the attacker recover raw input data, otherwise available only to the edge device?* Past work claimed the edge-cloud collaborative inference can provide better privacy protection, as the cloud only receives the intermediate values instead of the raw data [10].

Z. He and R. Lee are with Princeton University, Princeton, NJ, USA, 08540. Email: {zechengh, rblee}@princeton.edu.

T. Zhang is with Nanyang Technological University, Singapore, 639798. Email: tianwei.zhang@ntu.edu.sg.

An earlier version of this paper on only attacks [1] was presented at the 35th Annual Computer Security Applications Conference (ACSAC'19) and published in its Proceedings. https://dl.acm.org/doi/10.1145/3359789.3359824

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

We show that an untrusted cloud can still easily and accurately recover the sensitive data from the intermediate values without accessing the edge side model.

We design a set of novel attack techniques to achieve this goal under different settings. First, for a white-box attacker, we propose using Regularized Maximum Likelihood Estimation to recover the samples from the model parameters and intermediate values. Second, for a black-box attacker, we propose the Inverse Network attack to identify the reverse mapping from the intermediate outputs to inputs without the knowledge of model information. Third, we consider the most limited adversarial capability where the cloud has no knowledge of the target model, and is not allowed to query the model. Conducting privacy attacks under this setting is extremely difficult, and this threat model is rarely considered in past work. For these query-free attacks, we introduce a new method of Shadow Model Reconstruction to achieve this attack.

The second question we address in this paper is: *how can the edge devices mitigate privacy leakage from the untrusted cloud?* Past work adopted differential privacy to protect the inference data [13]. We show that this approach is impractical against our proposed attacks as it brings unacceptable performance degradation to the DL models. Instead, we propose two novel strategies that can better thwart the privacy attacks while still maintaining good model performance. The first one is *the dropout defense*: by deactivating random neurons during the inference, the adversary is not able to precisely generate the original images from the intermediate values. Our second defense is *privacy-aware DNN partitioning*: we comprehensively evaluate different factors that can affect the attack results, and propose some guidelines to partition the deep learning models for better privacy. We hope our findings can guide machine learning researchers and practitioners to design more secure collaborative inference systems.

The key contributions of this paper are:

- A systematic study of attacks and defenses for inference data privacy in edge-cloud collaborative machine learning systems.
- Three attack approaches to recover inference data under different settings.
- Two new defense approaches to prevent inference data leakage to the untrusted cloud.

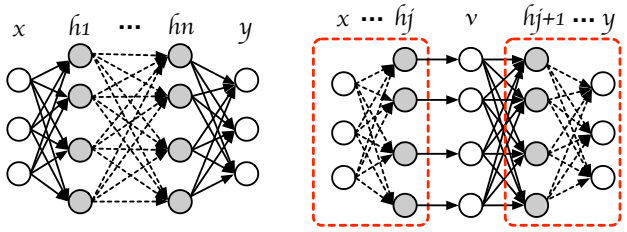
The rest of the paper is organized as follows: Section II presents the edge-cloud system model, threat model and experimental configurations. Section III describes attacks under white-box, black-box and query-free settings, including attack approaches, implementations and evaluation results. Section IV discusses possible mitigation solutions. We give related work in Section V and conclude in Section VI.

II. Preliminaries

A deep neural network (DNN) is a parameterized function $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ that maps an input tensor $x \in \mathcal{X}$ to an output tensor $y \in \mathcal{Y}$ (Figure 1a). It consists of an input layer, an output layer and a sequence of hidden layers between the input and output layers. Each layer is a collection of units called *neurons*, which are connected to other neurons in the previous layer and the next layer. Each connection between the neurons can transmit a signal to another neuron in the next layer. In this way, a neural network transforms the inputs through hidden layers to the outputs, by applying operations (e.g., a linear function or element-wise nonlinear activation function) in each layer.

A. System Model

In an edge-cloud collaborative inference system (Figure 1b), a DNN is partitioned into two parts: $f_\theta = f_{\theta 1} \circ f_{\theta 2}$. Each part contains several layers. The edge device hosts the first part $f_{\theta 1}$. It collects inference data from the environment, generates the intermediate value $v = f_{\theta 1}(x)$, and sends it to the cloud. The cloud hosts the second part of the model $f_{\theta 2}$. When receiving the intermediate value v from the edge device, it calculates the final output $y = f_{\theta 2}(v)$ and returns it to the edge device.



(a) DNN architecture (b) Collaborative inference

Fig. 1: DNN model (a) deployed in Collaborative edge-cloud system (b).

Determining a way to partition the DNN model is non-trivial. Different factors must be considered to identify the optimal strategy. (1) Latency: an optimal partition should give the fastest inference speed. The latency is determined by the inference time on the edge device, the cloud, as well as the network transmission time. The cloud can process the inference at a much faster speed. So it is preferable to move more DNN layers to the cloud. However, this can cause larger volumes of transmitted data, and longer network latency. So the performance of edge devices, cloud servers and network transmission must be balanced. (2) Power: an optimal partition should be energy-efficient. This is particularly important for edge devices which have limited power capabilities. The energy consumed by the edge device consists of the inference computation (determined by the number of layers) and network communication (determined by the size of transmitted data). Similar to latency optimization, energy consumption of these two parts needs to be balanced. (3) Memory size: when conducting inference, the device needs to load the entire DNN $f_{\theta 1}$ into the memory. Some edge devices are equipped with limited memory resources, and incapable of hosting too many network layers. This gives another constraint when selecting the optimal split point.

With these considerations, DNN partitioning is usually formulated as an optimization problem [8], [9], [10], [11], [12]. Figure 2 shows the comparisons of latency and energy consumption between edge-cloud, cloud-only and edge-only solutions (data are collected from [9]). We capture the results

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

from Figure 6 in Neurosurgeon [9]. In Neurosurgeon [9], a detailed study of latency and power consumption in a typical edge-cloud collaborative system was evaluated. An AlexNet model is deployed between a mobile device and cloud connected by WiFi. We observe that with an optimal split point, an edge-cloud system can achieve lower latency and energy than a cloud-only or an edge-only system: by offloading some DNN layers to the cloud, the processing time and energy consumed on the device is less than the edge-only system. Meanwhile, as the size of the intermediate data is smaller than the original input, the latency and energy costs of network transmission in the edge-cloud system are also less than the cloud-only system.

In practice, most layers (including all fully-connected layers) are commonly offloaded to the cloud, while the edge device only computes a small number of convolutional layers for feature extraction, due to power and resource constraints [9]. This gives a chance for an untrusted cloud provider to steal sensitive inference input, which we will discuss below.

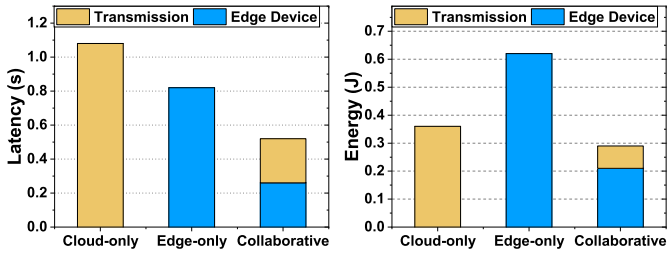


Fig. 2: Breakdown of inference latency (left) and energy consumption (right) in edge-cloud systems. Data are from [9].

B. Threat Model

We consider a collaborative inference system between the edge device $\mathbb{E}$ and cloud $\mathbb{C}$. The target model is split into two parts: $f_\theta = f_{\theta 2} \circ f_{\theta 1}$. $\mathbb{E}$ performs the first few layers $f_{\theta 1}$, while $\mathbb{C}$ performs the rest of the layers $f_{\theta 2}$. We consider $\mathbb{E}$ is trusted: when an input is fed into $f_{\theta 1}$, $\mathbb{E}$ correctly processes it and never leaks it to other parties. However, $\mathbb{C}$ is untrusted, attempting to steal the input. We consider confidentiality of an individual raw input when we use the term "data privacy" throughout this paper. Other forms of data privacy, e.g. membership or linkability, are not in our threat model.

We assume $\mathbb{C}$ strictly follows the collaborative inference protocol: receiving $v = f_{\theta 1}(x)$ from $\mathbb{E}$ and generating $y = f_{\theta 2}(v)$. $\mathbb{C}$ cannot compromise the inference process conducted by $\mathbb{E}$, and has no knowledge of the input x, nor any intermediate values inside $\mathbb{E}$, except v. We consider adversaries with different capabilities:

- *White-box*: $\mathbb{C}$ has the knowledge of the model at the edge side $f_{\theta 1}$, including its network structure and parameters.
- *Black-box*: $\mathbb{C}$ does not have knowledge of $f_{\theta 1}$, but is able to query the model $f_{\theta 1}$. The adversary does not need to know the exact training data, but he can collect the same type of samples as the training data. This assumption is reasonable in practice, e.g., the adversary can collect arbitrary face samples for a face recognition model or medical records for a diagnostic system.
- *Query-free*: $\mathbb{C}$ does not have knowledge of $f_{\theta 1}$, or the permission to query the model $f_{\theta 1}$. This type of attacks has the minimum attacker capability. Similar to the black-box attack, we assume the attacker can collect samples similar in type to the training data.

C. Notations and Experimental Configurations

We summarize our notations in Table I. We show detailed configurations of experiments in Table II.

Our attacks and defenses are generic and applicable to various datasets. In this paper, we demonstrate the attack results on the MNIST dataset and the defense results on MNIST and CIFAR10. More details on the attacks can be found in [1].

The first victim model we target is LeNet5. It consists of 2 convolutional layer blocks (each block has a convolutional layer, an activation layer and a pooling layer), 3 fully connected layers and 1 softmax layer. The model can be split at either the first convolutional layer, or the second convolutional layer after activation. These configurations are realistic in edge-cloud scenarios, as the heavy-computational layers (including all fully-connected layers) are offloaded to the cloud.

We follow the standard MNIST and CIFAR10 split for training and testing samples [14]. We set the learning rate to 10^{-3} and choose ADAM as our optimizer. The target model, all attack techniques and defense solutions are implemented with Pytorch 1.0.1. We run our experiments on a server with 1 Nvidia 1080Ti GPU and 2 Intel Xeon E5-2667 CPUs.

To quantify the effectiveness of attacks and defenses, we adopt two metrics, Peak Signal-to-Noise Ratio (PSNR) [15] and Structural Similarity Index (SSIM) [16]. Larger values of these two metrics indicate the recovered input is of higher quality, and more similar to the original one.

III. Attack Methodologies

A. White-box Attack

We start from the white-box setting, where the adversarial cloud knows the parameters of the initial layers $f_{\theta 1}$ on the edge device. Formally, the problem we consider is: *how can the adversary recover an input x_0, from the corresponding intermediate value $f_{\theta 1}(x_0)$, and the model parameters $\theta 1$?* We propose regularized Maximum Likelihood Estimation (rMLE) to solve this problem.

Regularized Maximum Likelihood Estimation. We treat the attack as an optimization problem: given $f_{\theta 1}(x_0)$, our goal is to find a generated sample x, that satisfies two requirements: (1) the intermediate output of this sample, $f_{\theta 1}(x)$, is similar to $f_{\theta 1}(x_0)$; (2) x is a natural sample, following the same distribution as other inference samples.

For requirement (1), we use the Euclidean Distance (ED) to measure the similarity between $f_{\theta 1}(x)$ and $f_{\theta 1}(x_0)$ (Eq. 1(a)). Note that $f_{\theta 1}(x)$ can be interpreted as the mapping from the input space (unobservable to the adversary) to the feature space

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

TABLE I: Table of Notations

Symbol	Description	Symbol	Description
$\mathbb{C}$	Untrusted cloud provider	S	Training set of f_θ
$\mathbb{E}$	Edge device	x_0	Original input
f_θ	Original DNN model	$x_{i,j}$	Pixel at position (i,j) in image x
$f_{\theta 1}$	Partial model at the edge side	ED($\cdot$)	Euclidean distance
$f_{\theta 2}$	Partial model at the cloud side	TV($\cdot$)	Total variation of an image
$f_{\theta 1}^{-1}$	Inversed network of $f_{\theta 1}$	$y \sim f_{\theta 2}(f_{\theta 1}(x))$	Output generated from the edge-cloud collaboration
σ	Standard deviation of Gaussian noise	β	Smoothness parameter of TV
b	Scale of Laplacian noise	λ	Balancing parameter of ED and TV
r	Dropout rate	m	Number of training samples

TABLE II: Experiment Configurations

Dataset	MNIST	CIFAR10
Target Model	LeNet-5 (2 conv + 3 fc)	6 conv + 2 fc CNN
Split points considered	• 1st conv layer (conv1) • 2nd conv layer after activation (ReLU2)	• 2nd conv layer (conv12) • 1st pooling layer (pool1) • 4th conv layer before and after activation (conv22 and ReLU22) • 2nd pooling layer (pool2)

(observable to the adversary). Then this Euclidean Distance represents the *posteriori* information from the adversary's intermediate-level observation. Our goal is to find the optimal sample x that minimizes this distance.

For requirement (2), we adopt the *Total Variation* [17] to represent the *prior* information of an input sample. The total variation of a 2D image x is defined in Equation 1(b), where $x_{i,j}$ represents the pixel at position (i, j). β is a parameter that controls the smoothness of the image. Larger β results in more piecewise-smoothed images. We set $\beta = 1.0$ throughout our experiments. Minimization of this metric can guarantee the generated image x is piece-wise smooth, i.e. avoiding drastic variations inside regions but allowing large changes along the region boundaries.

$$ED(x, x_0) = ||f_{\theta 1}(x) - f_{\theta 1}(x_0)||_2^2 \tag{1a}$$

$$TV(x) = \sum_{i,j}(|x_{i+1,j} - x_{i,j}|^2 + |x_{i,j+1} - x_{i,j}|^2)^{\beta/2} \tag{1b}$$

$$x^* = argmin_x \;\; ED(x, x_0) + \lambda TV(x) \tag{1c}$$

The total objective function of the model inversion problem is a combination of feature space similarity and input smoothness, as shown in Eq. 1c. In this equation, λ is a hyperparameter to balance the effects of the two terms. If the feature space $f_{\theta 1}(x)$ is far from the input space, i.e. a lot of network layers are computed on the trusted participant $\mathbb{E}$, a large λ is required because less posterior information about the input can be recovered from the feature space and the adversary needs to rely on the prior information. In contrast, if only a small number of layers are deployed on $\mathbb{E}$, then the adversary only needs to select a small λ. We set λ=0 when getting the inverse from layers before the first fully connected layer, and λ=0.1 when getting the inverse from layers after the first fully connected layer. We perform gradient descent (GD) to solve Eq. 1c and recover the image.

Evaluation. Figure 3 shows the white-box attack results. The first row shows the original inference samples, and the remaining rows are the recovered images when the split point is at different layers. We observe that the adversary can accurately recover the images with high fidelity, when the split point is either at the first (conv1) or last (ReLU2) convolutional layer. At the first split layer, PSNR is 39.69dB and SSIM is 1.00. At the last split layer, PSNR is 15.10dB and SSIM is 0.60. [1] This indicates that when the split point is at a deeper layer, the quality and similarity of recovered images become worse.

Fig. 3: Recovered inputs in white-box attacks

B. Black-box Attacks

Next, we consider the black-box setting, where the adversary does not have knowledge of the structure or parameters of $f_{\theta 1}$. We assume that the adversary can query the black-box model: he can send an arbitrary input x to $\mathbb{E}$, and observe the corresponding output $f_{\theta 1}(x)$.

Data privacy attacks under the black-box setting are more challenging, because without the knowledge of model parameters, the adversary cannot directly perform a gradient descent on $f_{\theta 1}$ to solve the optimization problem in Equation 1(c). One solution is to first recover the model structure and parameters by querying the model, and then recover the inference samples. The possibility of model re-construction has been demonstrated in [18], [19], [20].

We propose a more efficient approach, Inverse-Network, to directly identify the inversed mapping from output to input, without the model information. Our solution is easier to implement, and can recover inputs with higher fidelity.

Inverse-Network. Conceptually, the Inverse-Network is the approximated inverse function of $f_{\theta 1}$, trained with $v = f_{\theta 1}(x)$ as input, and x as output. The attack consists of three phases: ① generating a training set for the Inverse-Network; ② training the Inverse-Network; and ③ recovering the input sample by querying the Inverse-Network.

[1] In our experiments, we observe that PSNR>10dB or SSIM>0.3 are considered as good quality, because the inversed images are visually recognizable by the adversary.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

First, the adversary generates a bag of samples $X = (x_1, x_2, ..., x_m)$ of the same type as the training data to query the target system, and observes the corresponding intermediate outputs $V = (f_{\theta 1}(x_1), f_{\theta 1}(x_2), ..., f_{\theta 1}(x_m))$. Next, he can directly train an Inverse-Network $f_{\theta 1}^{-1}$ using V as the training input and X as the training output. We initialize the Inverse-Network with Xavier initialization [21], to avoid the neuron activations in the saturated or dead regions in the beginning. We leverage l_2 norm in the pixel space as the loss function (Eq. 2), and stochastic gradient descent (SGD) to train the Inverse-Network:

$$f_{\theta 1}^{-1} = argmin_g \frac{1}{m} \sum_{i=1}^{m} ||g(f_{\theta 1}(x_i)) - x_i)||^2 \quad (2)$$

where g is the Inverse-Network to be optimized. Note that the architecture of the Inverse-Network need not be related to the target model $f_{\theta 1}$. In our experiment, we use an entirely different network architecture.

Once the Inverse-Network $f_{\theta 1}^{-1}$ is obtained, the adversary can recover any inference sample from the intermediate layer output: $x = f_{\theta 1}^{-1}(v)$. This approach is more efficient than rMLE: (1) for each target sample, the adversary only needs to pass through the Inverse-Network once, while in rMLE, an iterative process is required to solve the optimization problem; (2) calculating the inversed input is parameter-free, while rMLE requires tuning the parameters (λ, β in Eq. 1).

Evaluation. Figure 4 shows the recovered images of MNIST. We can observe that the adversary can recover the input under the black-box setting with very high quality (PSNR is 40.72dB and 20.81dB for the two split points) and similarity (SSIM is 0.99 and 0.80 for the two points).

Fig. 4: Recovered inputs in black-box attacks

C. Query-Free Attacks

The Inverse-Network approach requires the adversary to be able to query the target model, and generate the data set for training f_{θ}^{-1}. In this section, we consider the query-free setting, where the adversary cannot query the model at the edge side, and does not know the model information. The basic idea is that the adversary first reconstructs a shadow model, which imitates the target model's behavior, and then uses rMLE over this shadow model to recover the input samples.

Shadow Model Reconstruction. The problem at the first step is: how can the adversary reconstruct a shadow model of the former model layers, $f_{\theta 1}$, with only the knowledge of the latter layers $f_{\theta 2}$ and the same type of training data as S? He cannot query the model with specified samples to get the intermediate values.

The key insight of our approach is that, if the shadow model is reconstructed as $f'_{\theta 1}$, it should be able to classify the input with high accuracy when combined with the later layers $f_{\theta 2}$:

$$y_i \sim f_{\theta 2}(f_{\theta 1}(x_i)) \sim f_{\theta 2}(f'_{\theta 1}(x_i)), \text{for } (x_i, y_i) \in S \quad (3)$$

Then the task of model reconstruction can be translated into minimizing the classification error of the composition of the two models: $f_{\theta 2}(f'_{\theta 1}(x_i))$ versus y_i. Eq. 4 shows the loss function for training the model, where m is the number of samples in S, $CrossEntropy$ is the cross-entropy loss. Equivalently, this means the training process of $f'_{\theta 1}$ is supervised at the output layer of $f_{\theta 2}$. Once the model $f'_{\theta 1}$ is reconstructed, the adversary can perform data recovery attacks using the rMLE technique in Section III-A.

$$f'_{\theta 1} = argmin_g \frac{1}{m} \sum_{i=1}^{m} CrossEntropy(f_{\theta 2}(g(x_i)), y_i) \quad (4)$$

$$CrossEntropy(\hat{y}, y) = -\sum_{c=1}^{C} y^c log(\hat{y}^c) \quad (5)$$

where C is the number of classes of the task.

There are two phases in our approach: ① offline shadow model reconstruction and ② online model inversion. The shadow model reconstruction only needs to be performed once. Then all the input samples can be recovered using the same shadow model, by only one inference for each input. In the shadow model reconstruction phase, the adversary can adopt the same type of samples as the training data. He may not know the original network structure $f_{\theta 1}$, but he can use an alternative one for the shadow model. We assume that both the target model and the shadow model are convolutional neural networks, but with different numbers of layers and filters, as well as filter sizes.

After the training set and network structure are determined, the adversary can adopt SGD to optimize the loss function of the composition of the two models. We choose the cross-entropy loss because it performs well on image classification tasks. Other loss functions can be leveraged, if the adversary aims to find inverses of the DNN for different tasks. Once the shadow model is obtained, the adversary can use rMLE to recover the inputs.

Evaluation We illustrate the recovered images under the query-free setting in Figure 5. The adversary can still recover the input images from conv1 and ReLU2 layers. The PSNRs for these two split points are 17.86dB and 8.03dB, while the SSIMs are 0.64 and 0.38, respectively. The quality of the images is relatively lower than the ones in the white-box or black-box setting, indicating the query-free attacks are harder to achieve. This is straightforward, as the adversary now has smaller capabilities. Also, more layers on the edge device can also increase the difficulty of image recovery.

IV. Defense Methodologies

Given the severity of inference privacy attacks in edge-cloud collaborative systems, we aim to explore defense methods in this section. We first empirically evaluate one common method

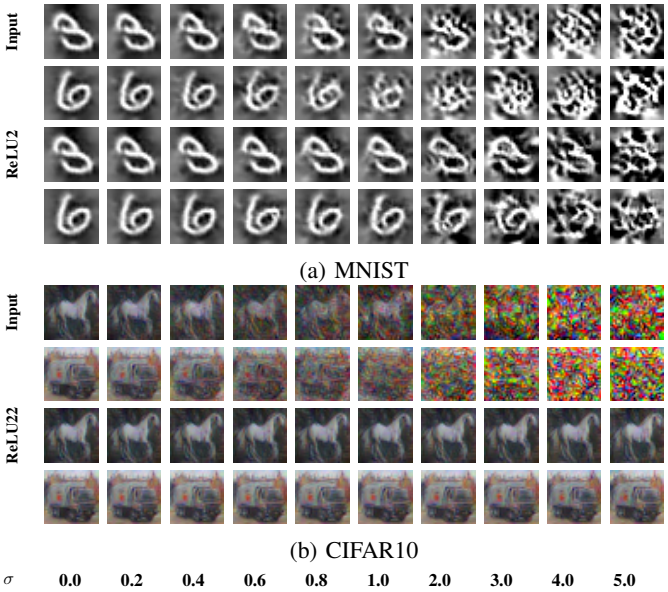
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

Fig. 5: Recovered input in query-free attacks

proposed in past work (though it did not specifically target the edge-cloud privacy attacks), viz., noise obfruscation. We show its ineffectiveness in defeating our inference data privacy attacks. Then we introduce two new strategies that can better prevent the privacy leakage with small impact to the system's performance and functionalities.

A. Obfuscation with Random Noise

Differential privacy has been proposed to protect model inference [22], [23] through adding random noise to the input. In the edge-cloud scenario, we can either add noise to the original input: $v = f_{\theta 1}(x + \epsilon)$, or add noise directly to the intermediate value before sending it to the untrusted cloud $\mathbb{C}$: $v = f_{\theta 1}(x) + \epsilon$. There is a trade-off between usability and privacy: as a higher level of noise is added, the model accuracy will drop. Whether this trade-off can be balanced is critical for the effectiveness of this approach. Below we measure the attack effects as well as the model accuracy using noise obfuscation.

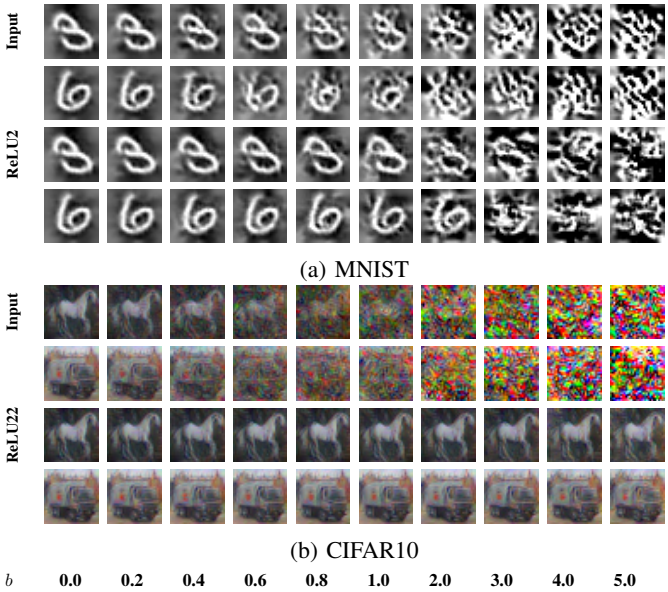


(a) MNIST

(b) CIFAR10

Fig. 6: Examples of adding Gaussian noise to defend against white-box attack on MNIST (ReLU2) and CIFAR10 (ReLU22). σ is the standard deviation of the Gaussian noise.

We consider Gaussian and Laplacian noise in our experiments. Figures 6a, 6b (Gaussian) and 7a, 7b (Laplacian) visually show the recovered images on the MNIST and CIFAR10 datasets, when we add different levels of noise to the input (first two rows in each figure) or the intermediate layer output (last two rows). We observe that adding enough noise can indeed provide better privacy and decrease the quality of recovered images. Besides, noise at the original input is more effective than noise at the intermediate layer.



(a) MNIST

(b) CIFAR10

Fig. 7: Examples of adding Laplacian noise to defend against white-box attack on MNIST (ReLU2) and CIFAR10 (ReLU22). b is the standard deviation of the Laplacian noise.

We provide a quantitative analysis of model accuracy (usability, y-axis) and inversed image quality (privacy, x-axis) in Figures 9 and 10 on MNIST and CIFAR10 dataset, repsectively. The Gaussian and Laplacian noise are represented as blue and orange lines, respectively. Adding noise to the input and intermediate layer are represented as solid and dotted lines, respectively. The top-left region of the graph is the best. When fixing the recovered image quality (SSIM or PSNR), the model accuracy drops more if the noise is added to the input (blue and orange solid lines) than to the intermediate layer (blue and orange dotted lines). This is consistent with the visual observations in Figure 6a and 7a. Different characteristics of noise distributions, e.g. Gaussian or Laplacian, do not show significant difference in model accuracy.

On the MNIST dataset (Figure 9), to maintain a good model accuracy (i.e., >95%), the noise level must be restricted to σ<0.8 and b<0.5. At this level, the attacker is still able to recover images with high quality (SSIM>0.4 and PSNR>8.5 dB). Similar results are shown on CIFAR10 dataset in Figure 10. While recent work [13], [6] proposed special algorithms for designing noise to protect inference data privacy, they still may not work for our new attacks, or need extra special training of the noise generator. Hence, we propose new defense methods below that are not based on adding noise, and are more practical in that they protect the inference data privacy with much smaller performance degradation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

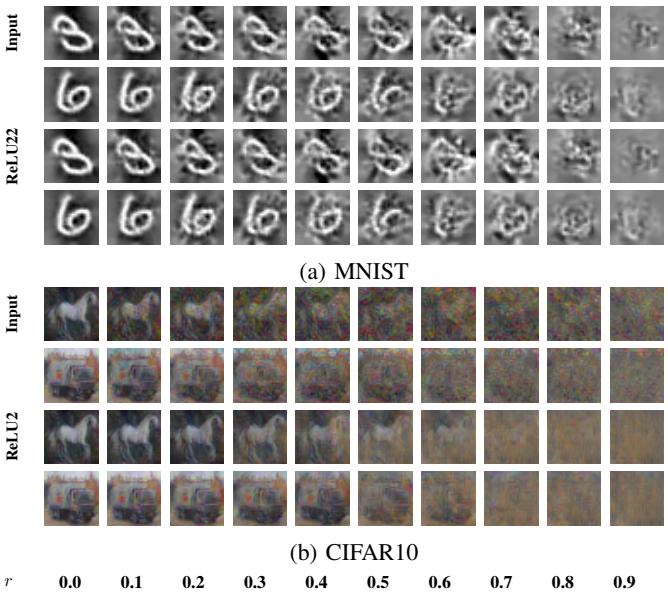
B. Dropout Defense

Since noise obfuscation may not be secure, we propose another randomization-based solution, dropout, to defeat the proposed attacks. Dropout deactivates random neurons in one layer by setting their output to zero. Formally, it calculates:

$$f_i^{dropout}(x) = f(x) \otimes M \tag{6}$$

where M is a mask, where each element of M is randomly assigned a value of 0 with probability r and a value of 1 with probability 1-r. $\otimes$ denotes element-wise multiplication. Intuitively, dropout leverages the redundancy feature of neural networks [24], such that removing partial information in the inference does not degrade the model performance but obfuscates the input data.

Similar to noise obfuscation, dropout can also be applied to the input or the intermediate layer output. We show examples of the images recovered from layer ReLU2 (MNIST) in Figure 8a. The top two rows represent the effect of dropout on input, while the bottom two rows represent that on intermediate output. We observe that increasing the dropout rate r decreases the quality of inversed images. No useful information can be obtained by the attacker when r reaches 0.6. We show reversed images from ReLU22 (CIFAR10) in Figure 8b. Similarly, no useful information can be obtained when r reaches 0.6.



(a) MNIST

(b) CIFAR10

Fig. 8: Examples of dropout to defend against white-box attack on MNIST (ReLU2) and CIFAR10 (ReLU22). r is the dropout ratio.

We further measure the usability-privacy trade-off of dropout, and compare it with the noise obfuscation approach (Figure 9 on MNIST and Figure 10 on CIFAR10). Higher accuracy represents better usability, while smaller SSIM and PSNR represent better privacy. Lines that are closer to the top left region have a better trade-off. We observe that dropout (green lines) significantly outperforms all the noise obfuscation solutions (blue and orange lines). This is because dropout leverages DNN model redundancy to hide partial information and maintain model accuracy, while adding random noise introduces obfuscation on all neurons which degrade model accuracy. Besides, dropout on the intermediate layer (green dotted line) is slightly better than dropout on the input (green solid line): it can fully protect the inference data privacy (SSIM<0.25) with accuracy>95%. On CIFAR10 dataset, dropout on the intermediate layer significantly overperforms the other approaches, fully protecting inference data privacy (SSIM<0.25) with <0.8% drop in accuracy.

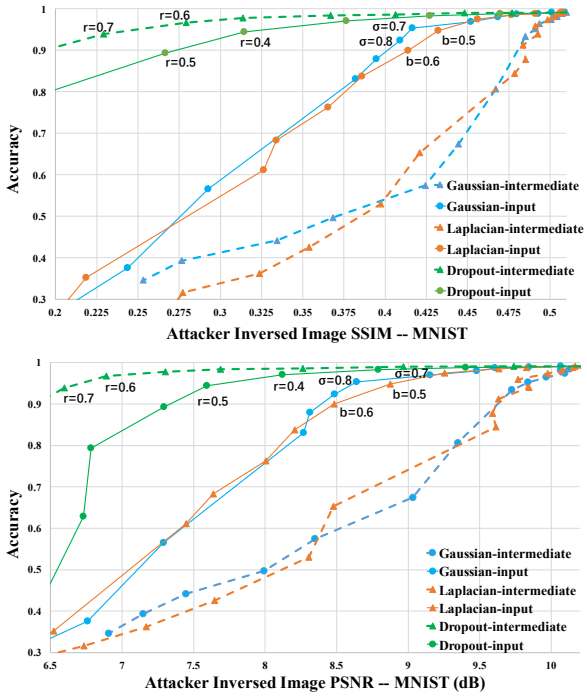


Fig. 9: Model accuracy versus the SSIM (top) and PSNR (bottom) of inversed images on MNIST dataset.

To fully evaluate the effectiveness of this dropout mechanism, we consider splitting the model at different layers on the MNIST dataset. We conduct dropout on the intermediate layer (i.e., split layer) since it is better than that on the input. Figure 11 shows the recovered images from shallow to deep layers: ReLU1, pool1, conv2, ReLU2, pool2. We observe that, as the split layer becomes deeper, a smaller dropout rate is sufficient to prevent privacy leakage. For example, to fully obfuscate the input, r can be set as 0.9 when the model is split at ReLU1 layer (first row), and 0.2 when the model is split at pool2 layer (last row). This can be better illustrated in the usability-privacy curves in Figure 13: dropout on deeper layers is more effective (closer to top left regions) than that on shallow layers. For both SSIM and PSNR, we have from worse to better: pool1(blue), then conv2 (green), then ReLU2 (orange), then pool2 (grey). There is only one exception: ReLU1 layer, which does better than expected. It is the best for SSIM and better than conv2 for PSNR. One possible reason is that the recovered image maintains visually recognizable structure but degrades illumination in ReLU1 layer, which contributes more significantly to PSNR and SSIM than human recognition.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

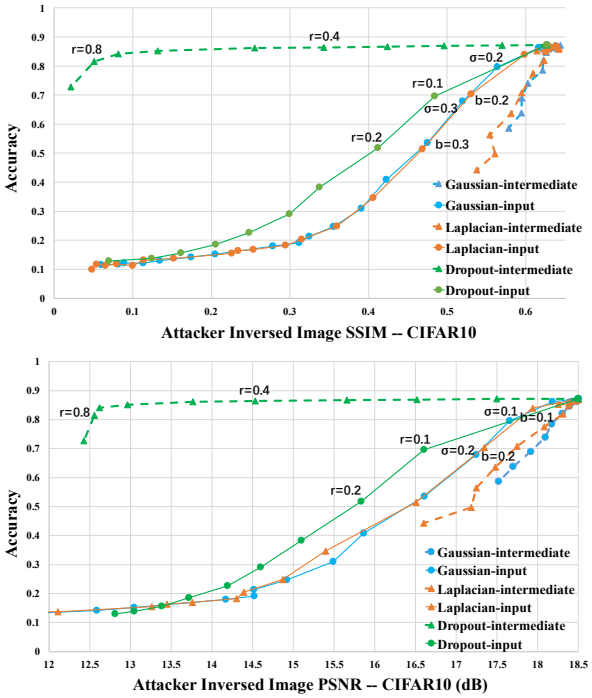


Fig. 10: Model accuracy versus the SSIM (top) and PSNR (bottom) of inversed images on CIFAR10 dataset.

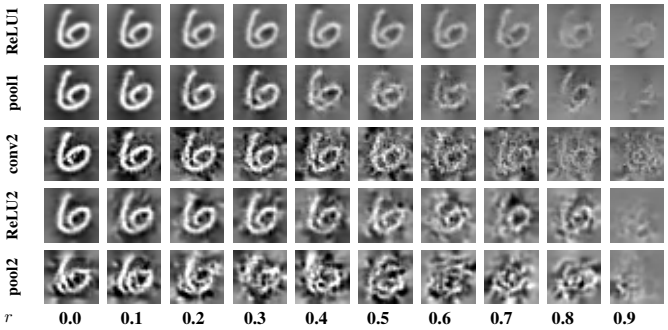


Fig. 11: Dropout as a defense against attacks at different layers on MNIST dataset. Rows from top to bottom: ReLU1, pool1, conv2, ReLU2, pool2.

C. Privacy-aware DNN partitioning

Section III shows that different split points yield different attack effects. This observation leads to another possible defense strategy: privacy-aware model partitioning. We raise an important question: *how to split the neural network in the collaborative system, to make the inference data more secure?* We use the query-free attack as an example to explore this question. We select the split point at each layer, and perform inference privacy attacks. Figures 14 and 15 show the recovered images, and PSNR/SSIM metrics respectively.

Generally, we observe that the quality of recovered images decreases when the split layer becomes deeper. This is straightforward as the relationship between input and output becomes more complicated and harder to revert when there are more layers. Besides, we also observe that the image quality drops significantly, both qualitatively (Figure 14) and

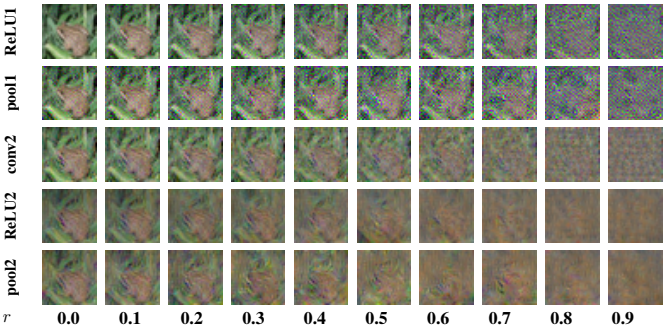


Fig. 12: Dropout as a defense against attacks at different layers on CIFAR10 dataset. Rows from top to bottom: ReLU12, pool1, conv22, ReLU22, pool2.

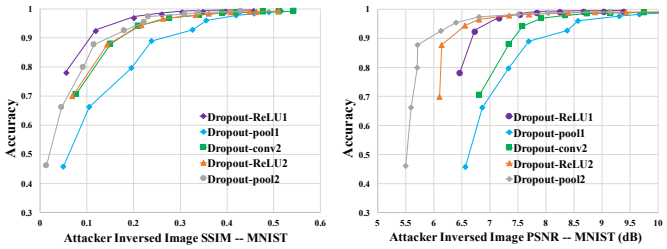


Fig. 13: Model accuracy versus the SSIM (left) and PSNR (right) of inversed image.

quantitatively (Figure 15), on the fully-connected layer (fc1), indicating that model inversion with fully-connected layers is much harder than that with convolutional layers. The reason is that a convolutional layer only operates on local elements (the locality depends on the kernel size), while a fully-connected layer entirely mixes up the patterns from the previous layer. Besides, the number of output neurons in a fully-connected layer is typically much smaller than input neurons. So it is relatively harder to find the reversed relationship from the output of the fully-connected layer to the input.

Privacy-aware partitioning strategy: When selecting the split point in a collaborative inference system, privacy should also be considered, in addition to latency and power constraints. We recommend placing at least one fully-connected layer on the edge device to hide the information of sensitive input samples.

V. Related Work

A. Machine Learning Privacy Attacks

Training data privacy attacks. There are different types of privacy attacks against the training data. The first type are *property inference attacks*, which try to infer some properties of the training data from the model parameters. Attacks were demonstrated in traditional machine learning classifiers [25] and fully-connected neural networks [26].

A special case of property inference attacks are *membership inference attacks*, which infer whether one individual sample is included in the training set. This attack was first presented in [27]. The following work explored the feasibility of attacks with different adversary's capabilities [28], model features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.3022358, IEEE Internet of Things Journal

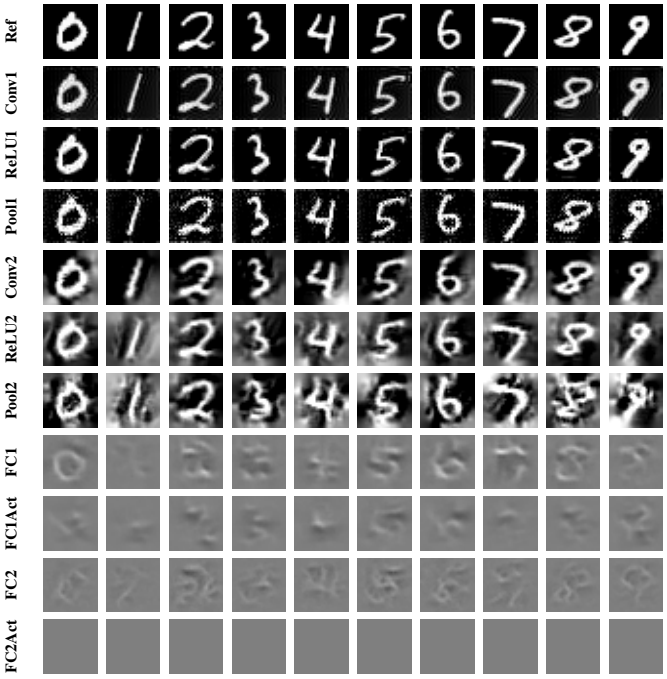


Fig. 14: Recovered images in query-free attacks.

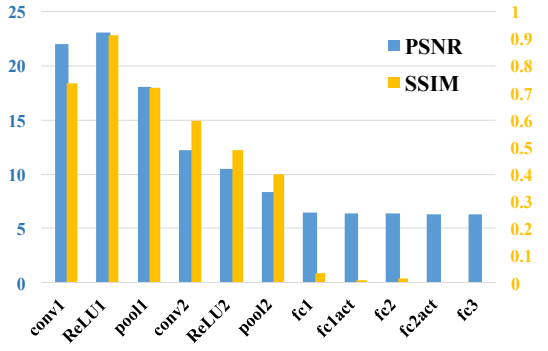


Fig. 15: PSNR and SSIM in query-free attacks.

[29], [30], in Generative Adversarial Networks [31], [32], and collaborative training systems [33].

The second type of attacks against the training data's privacy are *model inversion attacks* [34]: given a machine learning model, and part of the training samples' features, the adversary can recover the rest of the features of the samples. Advanced model inversion attacks were designed to recover images from deep neural networks in single-party systems [35], and collaborative learning systems [36].

The third type are *model encoding attacks* [37]: the adversary with direct access to the training data can encode the sensitive data into the model for a receiver entity to retrieve.

Model privacy attacks. The adversary attempts to steal the model parameters [18], hyperparameters [20] or structures [19], [38], via prediction APIs, memory side channels, etc.

Inference data privacy attacks. Closer to our study is the work [39], which trains an inverse network on the output probability distribution to get the inversed inference data. However, they only consider the model inversion attack from the softmax layer in the black-box scenario. We show that the attacker can successfully inverse the model from different layers, even in a stricter query-free scenario. We also provide defense strategies which are not discussed in their paper. Wei et al. [40] adopted a power side channel to recover inference data. However, this attack required the adversary to compromise the victim device for side-channel information collection, and it could only recover simple images (single pixel). Our work can recover any arbitrary complex data without access to, or knowledge of, the victim's device and computation.

B. Machine Learning Privacy Solutions

Enhancing the algorithms. Distributed training was introduced to protect the training data [41], [42], as different participants can use their own data for model training. The SGX security enclaves in Intel processors were used to protect the training tasks against privileged adversaries [43], [44]. Cao et al. [45] proposed a methodology to remove the effects of sensitive training samples on the models. Abadi et al. [46] applied differential privacy to add noise in the stochastic gradient descent process to eliminate the parameters' dependency on the training data.

Enhancing the training dataset. Bost et al. [47] proposed to encrypt the data before feeding them into the training algorithm. They designed machine learning operators which can operate on the encrypted data. Zhang et al. [48] showed that adding noise to the training dataset is effective in protecting training data privacy. Generating artificial data [49], [50], [51] has been proposed for training DNN models while removing sensitive information from the original data.

Obfruscating the inference input. Differential privacy has been proposed to protect model inference [22], [23] through adding random noise to the input. We show that just adding noise cannot defend against our attacks, and hence we also propose two defenses that may be more practical for our attacks in this paper. Recent work [6] proposed to add specially designed noise and provided a theoretical analysis on the input data privacy leakage. However, it did not consider the model inversion attacks that we propose and requires extra training of the noise generator.

Homomorphic encryption. This allows the inference application on the untrusted participant to directly perform DNN computations on encrypted input [52], [53], so the sensitive information will not be leaked. A drawback of homomorphic encryption is that it suffers from huge inefficiency and is not applicable for all DNN operations.

VI. Conclusions

In this paper, we explore the inference data privacy threats in edge-cloud collaborative systems. We discover that, an untrusted cloud can easily recover the inference samples from intermediate values. We propose a set of new attack techniques to compromise the inference data privacy under different attack settings. We demonstrate that the adversary can successfully and reliably recover the inputs with very few prerequisites.

We also propose several methods to protect the inference data privacy for edge computing. Previous work all focus on the performance, efficiency and functionalities of Artificial Intelligence of Things, while ignoring privacy. We hope that this study can raise awareness about the importance of inference data privacy protection in edge-cloud systems, and encourage the balancing of privacy protection with usability when designing or implementing such systems.

ACKNOWLEDGMENT

This work is supported in part by NSF STARSS grant 1526493 and a Qualcomm faculty award for Prof. Lee.

REFERENCES

[1] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 148–162.

[2] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," *Multimedia tools and applications*, vol. 76, no. 4, pp. 5817–5832, 2017.

[3] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 858–862.

[4] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[5] L. Xiao, Y. Li, X. Huang, and X. Du, "Cloud-based malware detection game for mobile devices with offloading," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2742–2750, 2017.

[6] F. Mireshghallah, M. Taram, P. Ramrakhyani, A. Jalali, D. Tullsen, and H. Esmaeilzadeh, "Shredder: Learning noise distributions to protect inference privacy," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 3–18.

[7] Z. He, T. Zhang, and R. Lee, "Sensitive-sample fingerprinting of deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4729–4737.

[8] J. Hauswald, T. Manville, Q. Zheng, R. Dreslinski, C. Chakrabarti, and T. Mudge, "A hybrid approach to offloading mobile image classification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 8375–8379.

[9] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *Acm Sigplan Notices*, vol. 52, no. 4, pp. 615–629, 2017.

[10] S. Teerapittayanon, B. McDanel, and H. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *IEEE International Conference on Distributed Computing Systems*, 2017.

[11] J. H. Ko, T. Na, M. F. Amir, and S. Mukhopadhyay, "Edge-host partitioning of deep neural networks with feature space encoding for resource-constrained internet-of-things platforms," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018.

[12] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, "Jointdnn: an efficient training and inference engine for intelligent mobile cloud computing services," *arXiv preprint arXiv:1801.08618*, 2018.

[13] F. Mireshghallah, M. Taram, A. Jalali, A. T. Elthakeb, D. Tullsen, and H. Esmaeilzadeh, "A principled approach to learning stochastic representations for privacy in deep neural inference," *arXiv preprint arXiv:2003.12154*, 2020.

[14] https://pytorch.org/docs/0.4.0/torchvision/datasets.html, 2018.

[15] https://en.wikipedia.org/wiki/Peak-signal-to-noise-ratio, 2018.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[18] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis." in *USENIX Security Symposium*, 2016.

[19] S. J. Oh, M. Augustin, M. Fritz, and B. Schiele, "Towards reverse-engineering black-box neural networks," in *International Conference on Learning Representations*, 2018.

[20] B. Wang and N. Z. Gong, "Stealing hyperparameters in machine learning," in *IEEE Symposium on Security and Privacy*, 2018.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[23] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[24] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2857–2865.

[25] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, 2015.

[26] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *ACM conference on computer and communications security*, 2018, pp. 619–633.

[27] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy*, 2017.

[28] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed System Security Symposium*, 2018.

[29] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium*, 2018.

[30] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding membership inferences on well-generalized learning models," *arXiv preprint arXiv:1802.04889*, 2018.

[31] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: evaluating privacy leakage of generative models using generative adversarial networks," *arXiv preprint arXiv:1705.07663*, 2017.

[32] K. S. Liu, B. Li, and J. Gao, "Generative model: Membership attack, generalization and diversity," *arXiv preprint arXiv:1805.09898*, 2018.

[33] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *IEEE Symposium on Security and Privacy*, 2019.

[34] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing." in *USENIX Security Symposium*, 2014.

[35] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM Conference on Computer and Communications Security*, 2015.

[36] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *ACM Conference on Computer and Communications Security*, 2017.

[37] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *ACM Conference on Computer and Communications Security*, 2017.

[38] W. Hua, Z. Zhang, and G. E. Suh, "Reverse engineering convolutional neural networks through side-channel information leaks," in *ACM/ESDA/IEEE Design Automation Conference*, 2018.

[39] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 225–240.

[40] L. Wei, B. Luo, Y. Li, Y. Liu, and Q. Xu, "I know what you see: Power side-channel attack on convolutional neural network accelerators," in *Annual Computer Security Applications Conference*, 2018.

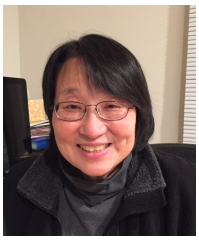
[41] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *ACM conference on computer and communications security*, 2015.

[42] J. Hamm, A. C. Champion, G. Chen, M. Belkin, and D. Xuan, "Crowd-ml: A privacy-preserving learning framework for a crowd of smart devices," in *IEEE International Conference on Distributed Computing Systems*, 2015.

[43] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa, "Oblivious multi-party machine learning on trusted processors." in *USENIX Security Symposium*, 2016.
[44] T. Hunt, C. Song, R. Shokri, V. Shmatikov, and E. Witchel, "Chiron: Privacy-preserving machine learning as a service," *arXiv preprint arXiv:1803.05961*, 2018.
[45] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *IEEE Symposium on Security and Privacy*, 2015.
[46] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM Conference on Computer and Communications Security*, 2016.
[47] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data." in *Network and Distributed System Security Symposium*, 2015.
[48] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," *arXiv preprint arXiv:1807.01860*, 2018.
[49] A. Triastcyn and B. Faltings, "Generating artificial data for private deep learning," *arXiv preprint arXiv:1803.03148*, 2018.
[50] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model (technical report)," *arXiv preprint arXiv:1801.01594*, 2018.
[51] H. Yin, P. Molchanov, Z. Li, J. M. Alvarez, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," *arXiv preprint arXiv:1912.08795*, 2019.
[52] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *International Conference on Machine Learning*, 2016, pp. 201–210.
[53] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "Gazelle: A low latency framework for secure neural network inference," in *27th USENIX Security Symposium*, 2018, pp. 1651–1669.

Zecheng He is a Ph.D candidate in the Department of Electrical Engineering at Princeton University. His research focuses on security and privacy in intelligent computer systems. He received his Bachelors degree from University of Science and Technology of China in 2015.

Tianwei Zhang is an assistant professor in School of Computer Science and Engineering, at Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture and distributed systems. He received his Bachelors degree at Peking University in 2011, and the Ph.D degree at Princeton University in 2017.



Ruby B. Lee is the Forrest G. Hamrick Professor of Engineering at Princeton University, and the Director of the Princeton Architecture Lab for Multimedia and Security (PALMS). Her current research is at the intersection of cyber security, computer architecture and deep learning. Her research includes improving security with deep learning, low cost deep learning processors and designing new architectures against attacks on microarchitecture like Spectre and Meltdown. Her past research includes architectures for secure processors and secure caches, and improving the security of smartphones and cloud computing servers. Prior to Princeton, Lee served as chief architect at Hewlett Packard in the computer systems division.