

PRIVADO: Practical and Secure DNN Inference with Enclaves

Karan Grover^{†*}, Shruti Tople^{†*}, Shweta Shinde[‡], Ranjita Bhagwan[†], and Ramchandran Ramjee[†]

[†]Microsoft Research [‡]UC Berkeley

Abstract

Cloud providers are extending support for trusted hardware primitives such as Intel SGX. Simultaneously, the field of deep learning is seeing enormous innovation as well as an increase in adoption. In this paper, we ask a timely question: “*Can third-party cloud services use Intel SGX enclaves to provide practical, yet secure DNN Inference-as-a-service?*” We first demonstrate that DNN models executing inside enclaves are vulnerable to access pattern based attacks. We show that by simply observing access patterns, an attacker can classify encrypted inputs with **97%** and **71%** attack accuracy for MNIST and CIFAR10 datasets on models trained to achieve 99% and 79% original accuracy respectively. This motivates the need for PRIVADO, a system we have designed for secure, easy-to-use, and performance efficient inference-as-a-service. PRIVADO is *input-oblivious*: it transforms any deep learning framework that is written in C/C++ to be free of input-dependent access patterns thus eliminating the leakage. PRIVADO is *fully-automated and has a low TCB*: with zero developer effort, given an ONNX description of a model, it generates compact and enclave-compatible code which can be deployed on an SGX cloud platform. PRIVADO incurs *low performance overhead*: we use PRIVADO with Torch framework and show its overhead to be **17.18%** on average on 11 different contemporary neural networks.

1 Introduction

In recent years, deep neural networks (DNNs) have revolutionized machine-learning tasks such as image classification, speech recognition, and language translation [1–3]. Today, the idea of applying DNNs to applications such as medical health prediction and financial modeling hold tremendous promise. Consider a medical health enterprise, Acme Corp., that has developed a state-of-the-art DNN-based model for identifying diseases from radiological images. Acme Corp. does not have its cloud infrastructure but wants to monetize this model by

making it available to hospitals worldwide using a third-party cloud provider. Acme wants to keep the model parameters (i.e., weights and biases) confidential since it is precious intellectual property. On the other hand, hospitals want to keep their radiology data secure given their privacy-sensitive nature. However, both Acme and its clients would still want to benefit from the cost-efficiencies of the cloud, and the functions it enables. Thus, a *secure deep learning inference service* is critical for addressing such scenarios.

Machine learning models hosted on the cloud are susceptible to several types of security and privacy issues [4–10]. In this work, we focus specifically on preserving the *confidentiality* of the model parameters and the inputs to the model from an attacker who can compromise the cloud software stack, or can gain physical access to the DRAM or the address bus. To protect code and data from compromised cloud software and malicious administrators, cloud providers are extending support for trusted hardware primitives such as Intel SGX [11, 12]. SGX supports hardware-isolated execution environments called *enclaves* that ensure that the user-code and data are inaccessible even to the cloud administrators. Nevertheless, we are still far from realizing an end-to-end DNN inference service using SGX enclaves due to the following three important challenges:

Security. Prior work has shown that enclave memory is susceptible to leakage via data access patterns [13–16]. However, it is not obvious that DNN model inference, that mostly uses matrix multiplications, is susceptible to such attacks. In Section 3, we motivate the need for data-oblivious DNN implementations within enclaves by demonstrating a *first-of-its-kind attack*. We show that an adversary can predict the input label with reasonably high accuracy (close to the original model accuracy) simply based on the observable access patterns during the execution of the model. Thus, we require the inference service to be *resistant to such memory address access pattern based attacks*.

Ease-of-use. Developing custom applications for enclaves is cumbersome. This is mainly because SGX does not provide system-call support, dynamic threading, etc. [17–19]. Thus,

*Lead authors ordered alphabetically.

the inference code has to be carefully partitioned into inside-enclave and outside-enclave code. Moreover, programming tools and debuggers that run within SGX are rudimentary [20, 21]. The alternative is to run entire applications that have been written for an insecure setting, unmodified, on entire operating systems and sandboxes that run within the enclave. Unfortunately, this results in a large trusted computing base (TCB) [17, 22] which is undesirable for security-sensitive applications. Ideally, the service should require *no custom programming* and yet, it should have a *low TCB*.

Performance. Finally, the solution can be practical only if the performance overheads are “acceptably” low.

Unfortunately, previous research on secure DNN inference using SGX does not solve these challenges entirely. Prior work has either concentrated on customized algorithms and code that can fix access-based attacks at the cost of ease-of-use [7], or on solutions that are easy to use but do not protect against access-based attacks and have a large TCB [9]. To make matters worse, these approaches have not been extensively evaluated on contemporary DNN models to ensure that performance overheads are indeed uniformly low. To this end, we present PRIVADO, a secure DNN inference service that simultaneously provides obliviousness, low TCB, ease-of-use, and low-performance overhead.¹ PRIVADO consists of two main components—PRIVADO-Converter and PRIVADO-Generator, that help us provide these properties. To our knowledge, PRIVADO is the first system that targets and solves all these challenges in a single system.

We first address the *security* challenge. The key observation that enables PRIVADO’s approach to data-obliviousness is that deep learning models predominantly have data-independent accesses and very few data-dependent accesses (Section 4). Our main contribution is the observation that deep learning models exhibit very specific and regular data access patterns that we call *assign-or-nothing* patterns. In programs with these patterns, every input-dependent conditional branch either assigns a value to a variable otherwise exits the branch. The presence of such a specific pattern allows us to completely automate the task of eliminating these memory accesses using simple yet efficient oblivious primitives, thereby making the inference process resistant to access-based attacks. Prior works have proposed manual modifications to neural network algorithms, thereby restricting the applicability of the solution to a handful of models [7].²

We present PRIVADO-Converter—a tool that automatically detects all such data-dependent access patterns in a given deep learning framework. It modifies them to become data-independent using any of the known oblivious constructs [7, 23, 24]. In our implementation, we select the CMOV instruction for this purpose. With PRIVADO-Converter, we transform the popular Torch framework to guarantee obliviousness [25].

¹PRIVADO is a Spanish word for *a confidential friend or a confidant*.

²Similar to the threat model in previous work [8–10], we do not provide protect the model structure such as the depth, neurons in each layer, etc.

This makes PRIVADO expressive to a large number of state-of-the-art models, as shown in our evaluation in Section 6. However, our observations about the access patterns in neural networks are independent of the framework and are applicable to deep learning algorithms in general.

To address the *ease-of-use* challenge, PRIVADO uses the PRIVADO-Generator which takes models represented in the popular Open Neural Network eXchange (ONNX) format [26] as input and automatically generates a minimal set of enclave-specific code and encrypted parameters for the model. PRIVADO does not require any custom programming and thus has zero developer effort. The server simply loads the auto-generated code for the model within an enclave where the model parameters are decrypted. To reduce the TCB size, the PRIVADO-generator only includes files required for building the model instead of the entire DL framework. This approach, we find, reduces the TCB size by **34.7%**. Finally, we address the *performance* challenge. We evaluate PRIVADO on **11** models used in real-world deep learning cloud services and show that PRIVADO has **17.18%** overhead on average for inference for MNIST, CIFAR10, and ImageNet image datasets.

In this paper, we make the following contributions:

- *Access Pattern Based Attack.* We demonstrate a concrete attack that predicts the output class of encrypted inputs to the neural network based on the access patterns observable in the intermediate layers. Our attack achieves **97%** and **71%** accuracy for predicting MNIST and CIFAR10 images on an MLP and LeNet model respectively.
- *End-to-end Design.* Starting with an ONNX model, we show that our system auto-generates data-oblivious code which runs seamlessly on SGX in the cloud with zero development effort for the cloud customer.
- *PRIVADO System.* We present PRIVADO, which incorporates the above-mentioned constructs into the Torch framework. PRIVADO converts Torch to replace data-dependent branches and it auto-generates minimal amount of Torch code from the ONNX model specification, thereby ensuring that the TCB size is small. Our evaluation shows that PRIVADO incurs an average performance overhead of **17.18%** on 11 models.

2 Problem

In this section, we first outline the problem setting for PRIVADO. We then describe our threat model. Finally, we outline the desirable properties that PRIVADO should achieve.

2.1 Secure Inference-as-a-service

Neural network algorithms operate in two phases—training and inference. The training step uses a labelled dataset to generate *model parameters* i.e., weights and biases such that the error between the true input class and the predicted class

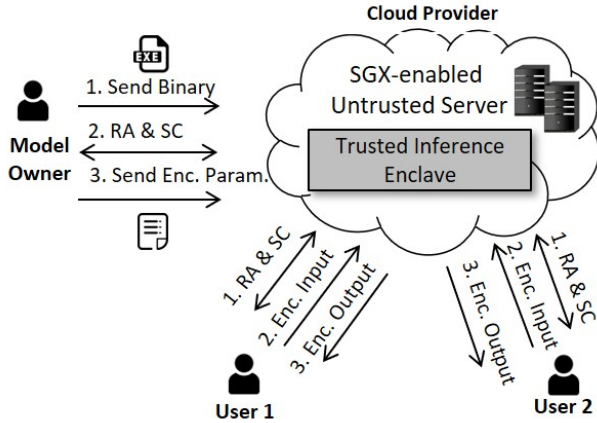


Figure 1: A secure inference-as-a-service setting. The model owner (1) sends the model binary, (2) performs remote attestation (RA) and establishes a secure channel (SC), and (3) sends the encrypted model parameters to the enclave. The model users (1) perform RA & SC, (2) send the encrypted inputs for inference, and (3) receive encrypted outputs.

is minimum. Once a network is trained, the inference phase uses the model parameters to accurately predict the output class for any given input. We assume that the training happens a priori in a trusted environment while the inference is offered as a cloud service to benefit other users or for monetary gains. To ensure confidentiality guarantees, such a cloud-based inference service should (a) protect the model parameters from the server and the users, and; (b) protect the users’ inputs and outputs from the server and other users of the service.

Figure 1 shows the entities involved in such an inference service: the *cloud provider*, the *model owner*, and multiple *model users*. We use the term model users and users interchangeably. The cloud provider supports trusted hardware primitives such as Intel SGX. SGX-enabled CPUs create isolated execution environments called *enclaves where all the code and data is protected from direct access via the untrusted software such as the OS as well as physical adversaries*. SGX allows remote attestation of the code running inside the enclaves to ensure its integrity [27].

Uploading & Instantiating the Model. A model owner has two components: (a) a model binary that contains the code corresponding to the model architecture; and (b) the model parameters (i.e., the weights and biases) generated after training the model architecture on a sensitive training dataset. The model binary / model architecture is public i.e., it is known to other entities such as the cloud provider or the users. On the other hand, the model parameters are the key intellectual property and hence are encrypted before sending to the server.

The model owner first uploads the model binary and attests the correctness of the loaded binary inside the enclave. After a successful remote attestation, the model owner establishes

a secure channel with the enclave that terminates inside the enclave [28]. Concretely, the model owner and the enclave perform a standard Diffie-Hellman key exchange to establish a shared secret key. Using this secure channel, the model owner provisions the model parameters to the enclave. When instantiating the model, the enclave decrypts these parameters using the shared key. This completes the process of hosting the inference service on the cloud.

Querying the Service. After the trusted inference enclave is set up, each user of this cloud inference service remotely attests the enclave to verify the correctness of the code executing inside it via standard methods as in other Intel SGX based cloud solutions [17, 29]. Once the attestation is complete, the user establishes a separate secure channel with the enclave to provision secrets. During this step, each user generates a distinct shared secret key with the enclave. To query the inference service, users encrypt their inputs with their respective secret keys and send them to the enclave. The enclave decrypts the input, runs the model owner’s binary on the user’s input, and encrypts the prediction output with the secret key corresponding to each user. Users receive the encrypted output label that can be decrypted only with their secret key. Throughout the process, the untrusted cloud software learns nothing about the model parameters (property of the model owner) or the input/predicted output (property of the model user). Further, the model owner learns nothing about the model users input/output and the model user learns nothing about the model parameters. As each user shares a separate key with the enclave, it is guaranteed that each user’s input remains confidential from all the other users.

2.2 Threat Model

In our threat model, we consider the cloud provider to be untrusted since an adversary can exploit bugs in the cloud software stack and get privileged access to the entire system. The only trusted entity is the SGX-enabled CPU processor available on the server. SGX guarantees that the hardware-isolated enclaves are inaccessible to the adversary. Enclaved execution guarantees that the encrypted content within the enclave gets decrypted only inside the processor package.

Although SGX prevents direct access to the enclave memory region, an adversary can learn significant information about the sensitive data via side-channels, specifically the access patterns. We consider software as well as a physical adversary who access the DRAM or can snoop over the address bus to learn the memory accesses and get the execution trace of the enclaved application. Also, the adversary can perform page fault and cache-channel attacks to learn the execution flow within the enclaves [13, 14, 30–37]. Moreover, the adversary can monitor the system call interface and observe access to the storage disks and trace network packets [16].

Assumptions. Similar to prior work in this domain [7–10],

we assume the adversary knows the hyper-parameters of the model such as the type of training data (e.g., images, text), model architecture, the number of layers, the number of neurons in each layer; these do not leak information about the sensitive inputs. We aim to provide secure inference service on a model without revealing the model parameters or the query input. Although leakage is possible via another side-channels such as timing, we do not address them in this work. However known solutions to mitigate timing channels can be used with PRIVADO [38–40]. Other types of attacks such as denial-of-service are not within the scope of this work. We assume that the cloud provider would always respond to the model users to maintain its reputation. Lastly, we assume that all the SGX guarantees are preserved i.e., there is no hardware backdoor present in the processor package and the secret keys are not compromised via any attack. We assume that remote attestation step and establishing of the secure channel are secure and are performed as per standard guidelines [41].

Out-of-scope: Privacy Attacks. A plethora of privacy-related attacks on machine learning models have been demonstrated such as model-inversion [5], model stealing [4], membership inference [6] and others. These attacks focus on compromising the privacy of either the training dataset or the model parameters. Model inversion and membership inference attacks target the training dataset while model stealing attacks aim to approximate the model parameters by observing the prediction score of the output. Although leakage via these attacks is an important concern, designing solutions to mitigate them is orthogonal to the main goal of PRIVADO. In PRIVADO, our primary focus is to ensure *confidentiality* of the original model parameters and the encrypted inputs of users while enabling a multi-user cloud-based inference service. PRIVADO does not ensure protecting the privacy of the data-points involved during the training step and hence is susceptible to these attacks. However, a model owner can use existing solutions such as differentially-private learning techniques or adversarial learning during the training phase before uploading the model to PRIVADO system [42–44].

2.3 Desirable Properties

PRIVADO aims to satisfy the following desirable design properties that are necessary to create a practical and secure inference service.

- *Input-Obliviousness:* We aim to achieve oblivious access patterns during the execution of the inference phase, thus preventing data leakage due to observable data-dependent memory address accesses.
- *Low-TCB:* The trusted computing base for any model should be the smallest possible subset of the entire deep learning framework to reduce the attack surface from bugs or vulnerabilities in the code.
- *Zero Developer Effort:* With PRIVADO, users need not

write custom SGX-specific code. Any machine learning scientist should be able to use PRIVADO out of the box.

- *Expressiveness:* We aim to support inference on state-of-the-art neural network models that contain complex combinations of linear and non-linear layers with parameters that range up to tens of millions.
- *Backward Compatibility:* We aim to support inference for models that have been trained in the past using any of the existing deep learning frameworks (e.g., Caffe, TensorFlow, PyTorch, CNTK).
- *Performance:* Lastly, we expect PRIVADO to have low performance overhead across several models trained on different input datasets.

3 Access Pattern Based Attack

We demonstrate that an adversary can indeed predict the labels of encrypted input data by merely observing access patterns of a DNN. Our high attack accuracies motivate the need to eliminate the leakage via memory access in the secure cloud-based DNN inference.

3.1 Target Model and Insight

We describe the attack strategy and perform the attack on two network models trained on MNIST and CIFAR10 datasets respectively. Our attack predicts the labels for encrypted inputs with **97%** and **71%** accuracy, respectively.

Target Model Architecture. We demonstrate our attack on two target deep neural network models trained using the Torch framework [25]. The first model is a multi-layered perceptron (MLP) model with 4 layers trained on the MNIST dataset [45]. MNIST is a collection of 32×32 -pixel black and white images of hand-written digits (0-9).³ We have trained the model to achieve 99% accuracy. Next, we train a LeNet model which is a convolutional neural network (CNN) with 3 convolution layer followed by 2 fully-connected layers. The model is trained to achieve 79% prediction accuracy on CIFAR10 dataset [46]. The CIFAR10 dataset has a collection of colored images for 10 different classes. Each hidden layer in both the models is followed by a rectified linear activation function (ReLU) layer. ReLU is a commonly used activation function to train accurate models. We perform our attack based on memory accesses observed only at these ReLU layers. Lastly, note that the final layer of both these models is a linear layer followed by a softmax function that outputs the probabilities of various classes. Note that we select MNIST and CIFAR10 as our example benchmarks, however the attack is feasible on any other dataset (such as ImageNet) trained on any model with data-dependent layers in the network.

³We used the Torch framework that provides 32×32 -pixel images instead of the common 28×28 for MNIST dataset.

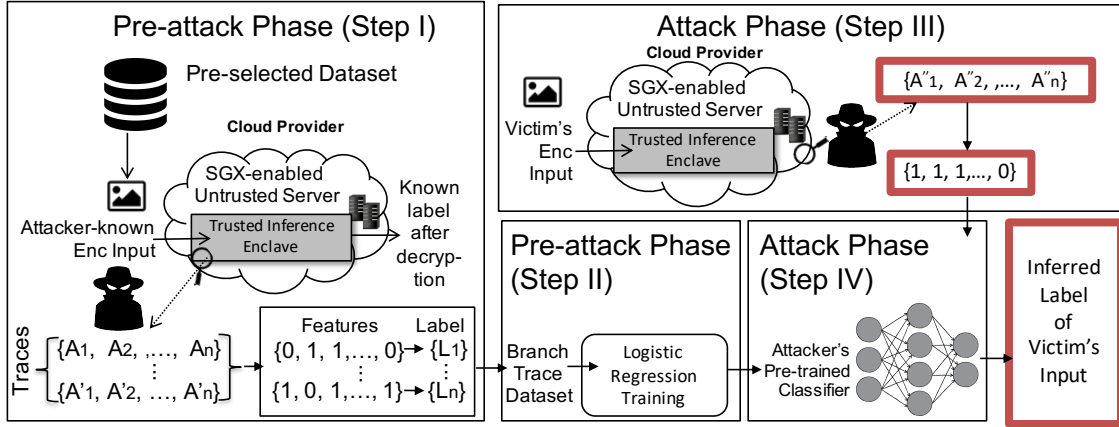


Figure 2: Attack steps where attacker: (I) executes its own dataset with known labels, collects traces, and extracts features; (II) uses the branch trace dataset (features, labels) for training a simple model; (III) observes the trace at runtime when the victim is executing inference on its own image inside the enclave, processes the trace to extract the features; (IV) passes the features to the pre-trained model to guess the label of victim’s input.

Leakage-prone Layer. The ReLU function is represented as $\max(0, x)$ which is commonly implemented using a conditional branch in several ML frameworks such as PyTorch⁴, Caffe⁵, and Tensorflow⁶. Listing 1 shows the code for the implementation of ReLU layer in the Torch library. The code is present in the `Threshold.c` file which provides a general threshold functionality and is not limited to the value zero.

```

1  if (*input_data <= threshold)
2      *input_data = val;
3  );

```

Listing 1: Input-dependent code from `Threshold.c` in Torch.

The function simply activates the neurons with values greater than zero (or threshold) and deactivates others. Consequently, the memory access pattern differs for every input based on whether the branch is executed or not. Therefore, observing the access patterns at the ReLU layer, an attacker can distinguish between neurons that are activated, and neurons that are not. In our attack, we exploit the relation between the activated neurons and the class label for any given input, without explicitly learning the exact neuron activations.

3.2 Attack Details

Figure 2 describes the key steps of our attack. In summary, the attacker performs the following steps:

- **(Step I).** In the pre-attack phase, first, the adversary uses its own input dataset to collect the memory access pat-

⁴<https://github.com/torch/nn/blob/master/lib/THNN/generic/Threshold.c>

⁵https://github.com/BVLC/caffe/blob/master/src/caffe/layers/relu_layer.cpp

⁶https://github.com/tensorflow/tensorflow/blob/master/tensorflow/core/kernels/relu_op_functor.h

terns and their corresponding output class predicted by the trained target model.

- **(Step II).** The attacker feeds these memory patterns as input features to train a classifier with logistic regression.
- **(Step III).** During the attack phase, to learn the output labels of other users (the victim), the adversary collects the memory address trace during the inference execution of the victim’s input.
- **(Step IV).** Finally, the attacker uses its trained classifier to predict the output class of victim’s input using the observed memory accesses.

Observing Memory Access Patterns. Here, we describe one concrete example of a memory access pattern trace that the attacker (the untrusted OS in this case) can collect via the page-fault channel. Since the OS is in charge of allocating physical memory to the enclave, it can severely limit the number of pages the enclave gets (say 1 page) [13]. Every time the enclave accesses an address on a different page, it incurs a page-fault which is delivered directly to the OS. The page-fault information consists of the page number and not the offset within the page (last three bytes of the address). Figure 3 (b) shows the actual addresses being accessed inside the enclave and Figure 3 (c) shows the page faults visible to the OS when executing LeNet model inference over two different inputs with labels `airplane` and `horse` respectively. If the attacker uses some other channel or their combination, in the worst case, it can learn the entire address (page number as well as the offset within that page). We refer to any of these as the memory address trace for simplicity.

Extracting Input Features for the Classifier. To extract the input features from the observed memory address traces, the attacker can do a `diff` between multiple traces. In our example, Figure 3, we see that most of the memory accesses

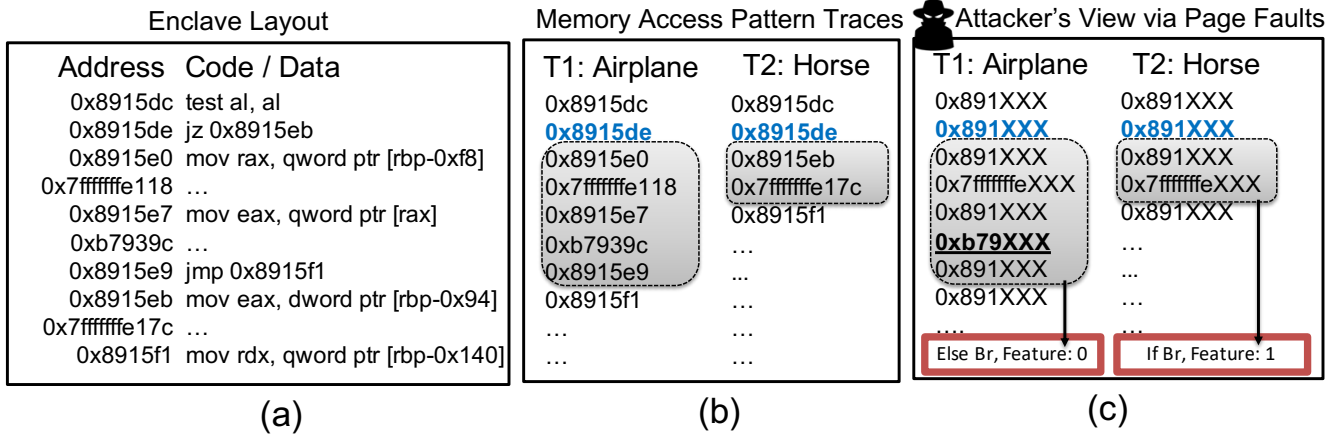


Figure 3: Enclave memory address layout and attacker’s observations. (a) The enclave address and code layout for ReLU when executing LeNet model. (b) Two different address traces exhibited when the enclave is executing LeNet model for inputs with labels airplane and horse respectively. The highlighted part in the dotted gray box shows the input-dependent difference in the trace. The bold blue text shows the branching instruction. (c) Page faults observed by the attacker when it allocates only one physical page to the enclave. When the attacker sees a page-fault for the underlined and bold page number it marks the feature value (indicated by box with thick border) to be 0 (if-branch not taken); otherwise it marks the value to be 1 (if-branch taken).

during the execution are deterministic. The location where the trace begins to differ, highlighted by bold and blue text in Figure 3 (b), indicates the execution of data-dependent branch condition in the code. Specifically, the address 0x8915de (highlighted in bold) belongs to the conditional statement that appears in both the traces. The address accessed after this location depends on whether the condition is true or false. More importantly, access to the address 0xb7939c causes a page-fault for page 0xb79 and reveals that the branch condition is not satisfied because of the if path is not executed. Whereas, when the enclave does not page-fault for this address, it reveals that the branch condition is satisfied.

Thus, by observing the difference in the traces of inputs from separate classes, the attacker not only learns the address locations for the data-dependent branch conditions in the code (i.e., ReLU in this model) but also whether the branch condition was satisfied or not. The attacker then extracts the features from the access pattern of each input. Specifically, for a given trace, it uses 1 to represent that the if branch executed and a 0 otherwise. The attacker does this for the entire trace which has multiple occurrences of ReLU because of a loop around the branch condition. At the end, the attacker has a vector of binary input features for each input image.

Creating Labelled Training Dataset. Our target model i.e., MLP or LeNet runs inside an enclave. The attacker acts as a legitimate user and establishes a secure channel with the enclave. Next, the attacker encrypts its input and queries the target model for inference. During the enclave execution, the attacker operating at the cloud provider records the memory address trace for the requested input via one of the

well-known Intel SGX cache side-channels [32, 34, 35] or controlled-channel attacks [13, 14, 31] (see next paragraph for details). The enclave responds with an encrypted output label to the attacker, who then uses its own key to decrypt the output label. The attacker repeats this process for various input images to collect memory address traces and corresponding labels. The attacker uses the above feature extraction method to generate binary input features corresponding to the memory access patterns. This becomes the labelled training dataset for the classifier.

Training Attacker’s Classifier. We use a simple logistic regression-based model for training the attacker’s classifier, where the number of input features for each input is equal to the total number of neurons in all the ReLU layers. Using input features proportional to the number of neurons may increase the computation cost for training the attacker classifier and grow significantly large for bigger models with millions of parameters. To reduce the number of input features in our attack, we use only the features that correspond to the last layer of the model. Since the hyper-parameters of the model are public, one can selectively generate only last n features from the memory address trace.⁷ Here, n corresponds to the number of neurons in the last layer. To understand if the attack accuracy improves with an increase in the number of features, we trained another classifier with input features generated from the last two ReLU layers of each of the two models. Further, we trained several instances of the attacker classifier for both the MNIST and CIFAR10 dataset by varying the number of training inputs. This allowed us to understand the

⁷We assume sequential execution trace of the model.

maximum number of inputs that the attacker needs to query the target model for building a highly accurate classifier. This is useful to estimate the cost of the attack as the inference service might charge the users for every query.

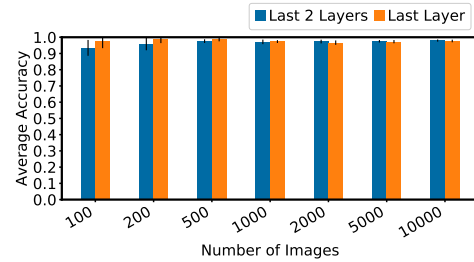
3.3 Attack Results

The attacker’s classifier achieves a 9-fold cross-validation accuracy of 97% and 71% for MNIST and CIFAR10 dataset respectively when trained on a dataset size of 10,000 inputs. This is the best-case accuracy when the attacker can query maximum number of inputs to the target model. We observe that the attacker accuracy is reasonably close to the original model’s accuracy i.e., 99% and 79% for MNIST and CIFAR10, respectively. Figure 4a and 4b gives the attack accuracy for predicting the labels of test dataset of MNIST and CIFAR10 images when the attacker classifier is trained over a range of training inputs.

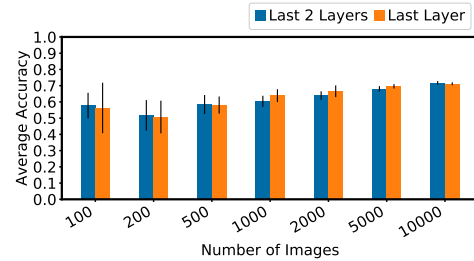
Note that the training features correspond to the memory access patterns of the last layer and the last two ReLU layers. This shows that the access patterns observed at the ReLU layer are critical and have the potential to reveal information about the sensitive inputs. These access patterns essentially capture the neuron activation information in these layers which is directly related to the output class. Thus, all the inputs that belong to the same class have similar address access pattern in the last layers. Note that, the attacker does not need to know the mapping of the neuron activation to the address patterns. The attacker classifier simply learns the memory address access pattern corresponding to each output class. This result can be explained from the understanding in machine learning theory that the last layer is the most representative of the input class (since this is what is used by the softmax layer to output class probabilities) [6]. Thus, an attacker who observes the access patterns only at the last layer is still capable of guessing the labels for the encrypted inputs with good accuracy.

We train the attacker classifier on access patterns with different number of query inputs. We observe that the attack accuracy is high for MNIST data even for smaller samples while it gradually increases for CIFAR10 with increasing in the training data size. This is due to the difference in the complexity of features in these two datasets. Thus, we observe that attacking a model trained on complex features would require larger number of queries to generate the inputs features for the attacker classifier. Consequently, when an honest user sends an encrypted image to the cloud, the adversary can apply the classifier on the observed access patterns and determine the input label of the encrypted image with high confidence. Therefore, it is critical that access patterns for DNN inference be made oblivious to prevent these attacks.

To conclude, we demonstrate very high attack accuracy in a worst-case attack scenario where the attacker has access to the entire memory address trace. However, similar attacks can be performed using either page-fault [13, 14, 31] or cache



(a) MNIST with a MLP model.



(b) CIFAR10 with LeNet model.

Figure 4: Accuracy results for access pattern based attack performed using a logistic regression model.

side-channel [32, 34, 35] to collect features corresponding to the branching of data-dependent conditional statement.

4 Sources of Leakage in DNNs

Leakage via access patterns is an important concern when using SGX for designing secure systems. Recently, researchers have proposed Oblivious RAM based solutions for SGX to eliminate leakage for arbitrary read-write patterns [24, 47]. However, these solutions incur significant performance overhead that makes them inefficient for a DNN inference service.

By customizing our solution to the specifics of how memory is accessed in DNN algorithms, we believe the overhead of making DNN inference data-oblivious can be significantly reduced. We make two key observations. First, the vast majority of computations in DNNs involve linear layers (fully connected or convolution layers) that exhibit only *deterministic* accesses, i.e., the memory access patterns do not vary based on the input. Second, certain types of DNN layers, such as ReLU or max-pool, exhibit input-dependent accesses, i.e., their memory access patterns can vary depending on the input (as we show in Section 3). However, even in layers that exhibit input-dependent accesses, the accesses are of a very specific type: in each branch either a given memory location is accessed or no memory locations are accessed. We call these *assign-or-nothing* patterns. This important observation with respect to the general class of deep learning algorithms allows us to build an efficient system that is resilient against access pattern based attacks. Further, this observation mainly allows us to automate the steps in PRIVADO without developer

effort, thus distinguishing it from prior work [7–10]. We discuss these observations in detail for layers that are commonly used in popular neural networks and provide supporting code having assign-or-nothing patterns in the Torch framework.

4.1 Linear and Batch Normalization Layers

In neural networks, such as multilayer perceptrons (MLPs) and convolutional networks (CNNs), the dominant part of the computation can be represented as *matrix-multiplication* between the weights and the inputs to each layer.⁸ In fact, over 90% of computation in many modern networks are attributed to the convolution operation [48]. Similarly, recurrent networks (RNNs, LSTMs) are also dominated by matrix multiplications [49]. Matrix multiplication involves performing the same operation, irrespective of input values. This makes their access patterns input-independent.

Batch Normalization [50] is another popular layer used in modern DNNs. Batch normalization, at inference time, adjusts the input value by the expected mean and variance of the population (computed during training). Thus, batch normalization does not perform any input dependent access.

4.2 Activation Layers

Activation layers are important in neural networks for capturing the non-linear relationship between the input and the output values. Several activation functions are used in these networks to improve the accuracy of the models. Sigmoid and tanh are the most basic activation functions which are computed as $\frac{1}{(1+e^{-x})}$ and $\frac{2}{(1+e^{-2x})} - 1$ respectively. As seen from the formulas, neither function incurs any input-dependent access patterns. While RNNs still use tanh activations, the standard choice in recent MLPs and CNNs have been the ReLU activation [51]. ReLU defined as $\max(0, x)$ uses an input-dependent conditional branch as we described in Section 3 and hence leaks information from its access patterns. The Torch framework offers 18 different activation functions out of which 15 exhibit assign-or-nothing patterns. They include ELU, Hardshrink, Hardtanh, LeakyReLU, LogSigmoid, PReLU, ReLU, ReLU6, RReLU, SELU, CELU, Softplus, Softshrink, and Threshold.

We consider HardTanh as an example. The functional definition of HardTanh is: $f(x) = 1$, if $x > 1$, $f(x) = -1$, if $x < -1$, $f(x) = x$, otherwise. Listing 2 shows the source code from the HardTanh.c file in the Torch framework. For each of the conditions, the branch only executes an assignment operation if true, otherwise, there is no assignment. Therefore, HardTanh exhibits the assign-or-nothing pattern. Further, observe that the condition is sequentially executed for all the inputs (or indices). The if condition is executed inside a for loop that

⁸Sometimes convolutions are computed using fast-fourier transforms but they are also input-independent.

iterates over all the inputs. All the loop bounds are public values such as the model hyper-parameters which do not leak information about the sensitive parameters and the user inputs. Observe that there are no nested branches that are conditioned on sensitive inputs (such as weights and bias). The inplace variable is public and its value can be decided a priori. We do not encounter accesses to memory locations with sensitive index. All the other activation functions exhibit a similar use of assign-or-nothing patterns in their implementation.

```

1  real* ptr_input = THTensor_(data)(input);
2  real* ptr_output = THTensor_(data)(output);
3  ptrdiff_t i;
4  ptrdiff_t n = THTensor_(nElement)(input);
5
6  if (inplace)
7 #pragma omp parallel for private(i)
8     for (i = 0; i < n; i++){
9         if (ptr_input[i] < min_val)
10            ptr_input[i] = min_val;
11        else if (ptr_input[i] > max_val)
12            ptr_input[i] = max_val;
13    }
14    else
15 #pragma omp parallel for private(i)
16        for (i = 0; i < n; i++){
17            if (ptr_input[i] < min_val)
18                ptr_output[i] = min_val;
19            else if (ptr_input[i] <= max_val)
20                ptr_output[i] = ptr_input[i];
21            else
22                ptr_output[i] = max_val;

```

Listing 2: Input-dependent code from HardTanh.c in Torch.

4.3 Pooling Layers

It is common to use pooling layers after convolution layers to reduce the output size at each layer. Two popular variants of pooling layer are max-pool and mean-pool. Typically, a filter of size 2×2 with a stride of 2 is applied to the input of this layer. The pooling function replaces each of the 2×2 region in the input either with a mean or max of those values, thus reducing the output size by 75%.

As the mean-pool function simply computes an average of the values from the previous layer, it exhibits a deterministic access pattern irrespective of the actual input to the model. However, the max-pool variant selects the maximum value in this 2×2 window and sets it as the output. Listing 3 shows the code that implements a max-pool operation.

```

1 void THNN_SpatialDilatedMaxPooling_updateOutput
2 (...) {
3     ...
4     if (val > maxval)
5         maxval = val;

```



```

6 ...
7 /* set output to local max */
8 *op = maxval;

```

Listing 3: Example of an input-dependent branch.

The input-dependent branch statement writes to the sensitive input location only if the condition is true. The loop terminating conditions are public values (2×2) that does not leak information about the input. Hence, similar to the ReLU function, the max-pool operation exhibits an assign-or-nothing pattern. The Torch framework offers 3 types of pooling functions `SpatialMaxPooling`, `TemporalMaxPooling`, and `VolumetricMaxPooling`. All of them exhibit assign-or-nothing patterns.

4.4 Softmax Layer

Most of the networks conclude with a softmax layer that calculates the probability for each of the output classes. The softmax layer performs a deterministic computation of calculating a value corresponding to each output class. The class that has the highest value is then returned as the predicted class. Although the softmax layer is oblivious in itself, computing the maximum value among all the classes requires an input-dependent branch condition to find the max value, similar to the max-pool function. For this, a sequential comparison is performed over all the output values to find the class with the maximum probability. Again, this falls into the category of assign-or-nothing patterns.

4.5 Summary

The above analysis of all the common layers used in neural network algorithms during inference show that all the sensitive input-dependent access patterns can be categorized into the assign-or-nothing pattern. Moreover, the dominant computation cost (90+%) is matrix-multiplications that are input independent. Thus, a custom data-oblivious solution for DNNs that addresses only the input dependent access pattern is likely to have low overheads. We do not consider the training of a model and therefore avoid operations that leak information during back-propagation.

5 PRIVADO Design

In this section, we present an overview and the design details of PRIVADO which consists of PRIVADO-Generator and PRIVADO-Converter that are trusted.⁹

5.1 Overview

The model owner generates an inference model for the PRIVADO system using two steps (Figure 5). In Step 1, the model

⁹The code is small and can be verified by inspection.

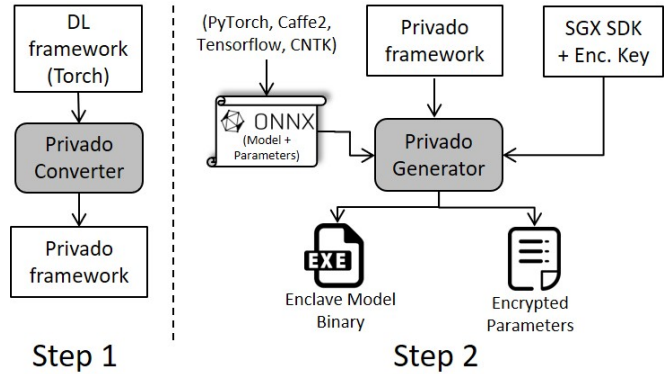


Figure 5: Overview of how the model owner generates the encrypted model and parameters using the PRIVADO-Converter and the PRIVADO-Generator.

owner uses PRIVADO-Converter to transform a deep learning framework to be input-oblivious. In our prototype, we have designed PRIVADO-Converter to transform the popular Torch library. This is a one-time process for any given framework. PRIVADO-Converter identifies all the assign-or-nothing patterns in the DNN framework and replaces them with the input-oblivious primitives. The converter is generic with regards to the underlying oblivious primitives. Any oblivious construction can be plugged into the PRIVADO-converter as long as the construction guarantees obliviousness for all the branch statements involving sensitive inputs. We select the CMOV-based solution that has been used to ensure obliviousness in previous work [7, 23, 24].

In Step 2, the model owner uses PRIVADO-Generator to generate the model binary. PRIVADO-Generator takes as input the ONNX representation of the model, the input-oblivious DNN framework that Step 1 generated, required SGX libraries, and an encryption key. It outputs an enclave-executable model binary and the encrypted model parameters. ONNX is an open neural network exchange format [26] that is supported by several existing deep learning frameworks (e.g., Caffe2, Tensorflow, CNTK, and others). Finally, the model owner uploads the model binary to the cloud provider to host it as an inference service, as described in Section 2.1.

5.2 PRIVADO-Converter

In this section, we outline how PRIVADO-Converter identifies input-dependent conditions and makes them oblivious.

Identifying Input-dependent Branches. DNN frameworks consist of libraries that the model uses to construct the final model binary. PRIVADO-Converter’s first step is to statically analyze the source code of these libraries and identify the branches. To do this, PRIVADO statically traverses the AST of each function in the library and reports conditional statements such as if-else, input-dependent loop guards, and ternary oper-

ations. PRIVADO then performs an inter-procedural data flow analysis to identify all the input-dependent variables [52]. Finally, PRIVADO-Converter collects all the variables that are involved in each conditional statement and selects only the ones that are input-dependent.

Next, PRIVADO-Converter categorizes the input-dependent conditional statements into *public* and *private* input-dependent branches. Specifically, we white-list all the branches that use purely public input, i.e., hyper-parameters such as network size, input size, etc. Then, PRIVADO-Converter performs taint analysis [53] with sensitive variables such as query input, weights, and biases marked as taint sources. PRIVADO’s analysis propagates the taint and identifies the program variables in the sinks i.e., the conditional statements of the deep learning framework that leak private information. Thus PRIVADO can check if any of the branches use private input-dependent variables. It marks all such branches that can potentially leak the user’s private data. Listing 3 shows one such branch in the Torch library implementation of the Max-pool layer, where the variables `val` and `maxval` used in the if-else condition on Line 4 are derived from the input. `val` and `maxval` are activation values of neurons that are provided as input to the pooling layer.

Our automated analysis provides us with a candidate list of private input-dependent branches, i.e., branches that may potentially leak inputs. We study a sampled set of the branches flagged by our analysis and observe that all of them adhere to the assign-or-nothing pattern. We do a best-effort source code analysis using a compiler pass to identify well-known branching constructs such as loops, if-else, and ternary operators. It is possible that our static analysis may miss some branches (e.g., inline assembly code). To this end, we further empirically cross-checked our branch analysis by comparing dynamic execution traces of the library for multiple inputs in our evaluation. Our results confirm that our best-effort analysis at least detects all the branches involved in the network implementations that we test PRIVADO on (See Section 6.6).

In the future, dynamic monitoring can ensure that PRIVADO never misses any branches that leak information, because it did not statically detect a branch in the library. Specifically, similar to binary-instrumentation techniques for CFI, an enclave-level mechanism can flag new branches whenever it observes that the input-dependent execution traces are about to deviate the execution trace [53]. Thus, our present static analysis can be combined with dynamic monitoring to ensure the completeness of our branch analysis.

Using CMOV for Obliviousness. We use CMOV in a way similar to previous work [23]. CMOV is an x86 and x86-64 instruction that accepts a condition, a source operand, and a destination operand. If the condition is satisfied, it moves the source operand to the destination. The important thing to note is that the CMOV instruction is oblivious when both the source and the destination operands are in registers: it does not cause any memory accesses, irrespective of the condition.

The key difference from previous work is that PRIVADO-Converter automatically modifies all occurrences of such branches in the DNN library with equivalent code that uses CMOV with registers. Moreover, using CMOV is just an implementation choice. Alternatively, we can use other oblivious primitives such as AND and XOR in PRIVADO-Converter. Let us consider the (simplified) branch code in Listing 4 that follows the assign-or-nothing pattern.

```
1 if (x < y)
2   x = y;
```

Listing 4: Simple example of assign-or-nothing pattern.

Listing 5 shows how PRIVADO-Converter replaces Listing 4 with a functionally equivalent oblivious code.

```
1 //if (x < y) x = y;
2 mov eax, x
3 mov ebx, y
4 cmp eax, ebx
5 cmovl eax, ebx
6 mov x, eax
```

Listing 5: Oblivious code by replacing < operator with `cmovl`

First, the code copies sensitive values `x` and `y` into registers `eax` and `ebx` respectively. Then, the CMOV instruction does a register-to-register copy based on the output of the comparison instruction. Note that the adversary cannot observe the register values in the secure CPU package. Therefore, using CMOV instruction makes memory access patterns deterministic and oblivious. This ensures that after our transformation, the library does not leak any information. PRIVADO-Converter employs different conditional move instructions (e.g., `CMOVL`, `CMOVZ`, `CMOVE`, `CMOVBE`, `FCMOVBE`) based on the specific instances of the branch patterns as well as the data types of the variables. Listing 6 shows the use of `FCMOVBE` instructions to replace the leaky branch in the max-pool example in Listing 3.

```
1 float temp;
2 asm volatile("fld %2 \n"
3             "fld %3 \n"
4             "fcomi \n"
5             "fcmovbe %%st(1), %%st \n"
6             "fstp %0 \n"
7             "fstp %1 \n"
8             : "=m" (maxval), "=m" (temp)
9             : "m" (val), "m" (maxval));
```

Listing 6: Using conditional move instructions to hide input-dependent branch in Listing 3.

Automated Transformation. PRIVADO-Converter uses the LLVM compiler for the source-to-source transformation of the input-dependent branches. Thus, we automate the transformation and ensure that the code uses CMOV. Our analysis

ensures that the transformed code preserves the intended functionality of the algorithm. Since the number of branches which are transformed is relatively small, we manually check that the transformed code is functionally equivalent to the original code. We also empirically verified this claim re-running the inference to ensure that the accuracies of the models are indeed preserved after PRIVADO-Converter’s transformation.

5.3 PRIVADO-Generator

The PRIVADO-Generator takes in an ONNX representation of a trained model and outputs an enclave-executable model binary and encrypted parameters. The ONNX format file contains the model configuration details such as the layers, neurons as well as the values for weights and biases.

Generating Enclave-specific Code. For a given model, PRIVADO-Generator generates two pieces of code: one executes inside the enclave and the other executes outside. It also generates an edl file which defines ecalls (entry calls to the enclave) and ocalls (exit calls from the enclave). Specifically, PRIVADO-Generator defines two ecalls: `initialize()` initializes the parameters of the model and `infer(image)` performs inference on encrypted images. It defines one ocall, `predict()`, which returns the model’s (encrypted) predicted class. PRIVADO-Generator generates enclave-bound functions by parsing the ONNX model and converting the ONNX operators to corresponding Torch function calls. For example, some ONNX operators have an equivalent function in Torch (ReLU’s equivalent is `THNN_FloatThreshold_updateOutput()`) while other operations, like grouped convolution, require composing multiple Torch functions in a for-loop. The non-enclave code deals with creating and initializing the enclave, waiting on a socket to receive client connections and encrypted images, calling the ecall to perform inference, and returning the predicted class to the client.

Reducing TCB. PRIVADO-Generator trims the Torch library to only the bare minimum set of files required to compile a given model. Once it identifies all layers in the ONNX model, it includes and compiles only the necessary Torch files from the math and the NN libraries. This step excludes irrelevant library code and thereby reduces the TCB.

Encrypting Parameters. The model weights and parameters are encrypted by the model owner. The owner shares the encryption key over a secure channel to the enclave that executes the model binary as described in Section 2.1.

6 Evaluation

In this section, we first provide a brief outline of the PRIVADO implementation. Next, we state our evaluation goals. Finally, we describe the results of our evaluation.

| Models | No. of Layers | No. of Param. | Dataset | Acc. (%) Top 1/ 5 |
|-------------|---------------|---------------|----------|-------------------|
| MLP* | 6 | 538K | MNIST | 99 |
| LeNet* | 12 | 62 K | Cifar-10 | 79 |
| VGG19 | 55 | 20.0 M | Cifar-10 | 92.6 |
| Wideresnet* | 93 | 36.5 M | Cifar-10 | 95.8 |
| Resnext29* | 102 | 34.5 M | Cifar-10 | 94.7 |
| Resnet110* | 552 | 1.70 M | Cifar-10 | 93.5 |
| AlexNet | 19 | 61.1 M | Imagenet | 80.3 |
| Squeezenet | 65 | 1.20 M | Imagenet | 80.3 |
| Resnet50 | 176 | 25.6 M | Imagenet | 93.6 |
| Inceptionv3 | 313 | 27.2 M | Imagenet | 93.9 |
| Densenet | 910 | 8.10 M | Imagenet | 93.7 |

Table 1: Models evaluated with PRIVADO. Columns 1 – 5 show the model name, the total number of layers, the number of parameters, the dataset and the accuracy respectively. * indicates we trained these models ourselves; rest of the models were taken from ONNX repository of pre-trained models [54]. We show Top 1 and Top 5 accuracies we observed in our inference for MNIST, Cifar-10, and ImageNet datasets.

6.1 Implementation

Our PRIVADO prototype is built on the Torch DNN framework. PRIVADO-Converter is implemented as an LLVM source-to-source transformation pass and has 1437 lines of C code. PRIVADO-Generator has 1700 lines of C++ code and uses the `sgx-crypto` library to implement the encryption-decryption functions in the enclave. We use Intel SGX SDKv.2 for Linux and compile using GCCv5.4 with `-O2` optimization flag. We select 11 state-of-the-art models for our experiments (Table 1).

6.2 Evaluation Goals

Our evaluation answers the following questions:

- *Ease-of-use.* How easy is it to use PRIVADO to generate secure versions of various state-of-the-art models?
- *Performance.* How much overhead does PRIVADO add, compared to baseline (insecure) inference?
- *Lowering TCB.* How much TCB does PRIVADO reduce?
- *Obliviousness.* Are the generated models oblivious?

6.3 Ease-of-use

In our experiments, we used PRIVADO-Generator on 11 state-of-the-art neural network models specified in ONNX. We select these models based on the differences in their depth (or the number of layers), parameter sizes, training dataset, and accuracy as shown in Table 1. PRIVADO-Generator successfully transformed all these trained models from their given ONNX format to SGX-enabled code within a few seconds. No custom-coding was required. We chose networks that

| Models | Execution time (in sec) | | | Overhead (in %) | Enclave mem size (in MB) | Normalized Page-Fault per sec | TCB Reduction (in %) |
|-------------|-------------------------|----------|------------|-----------------|--------------------------|-------------------------------|----------------------|
| | Baseline | SGX | SGX + CMOV | | | | |
| MLP | 0.0007 | 0.000569 | 0.000571 | -18.7 | 3.08 | 0 | 44.9 |
| LeNet | .001 | .0007 | 0.000726 | -27.4 | 0.624 | 0 | 34.4 |
| VGG19 | 0.62 | 0.61 | 0.61136 | -1.4 | 89.7 | 0 | 33.2 |
| Wideresnet | 12.3 | 14.2 | 14.2252 | 15.6 | 248.8 | 5080 | 33.4 |
| Resnet110 | 0.38 | 0.32 | 0.3242 | -14.6 | 81.3 | 0 | 33.4 |
| Resnext29 | 9.14 | 10.9 | 10.907 | 19.3 | 267.9 | 7302 | 33.4 |
| AlexNet | 0.99 | 1.88 | 1.885 | 90.4 | 258.0 | 65245 | 33.8 |
| Squeezenet | 0.49 | 0.38 | 0.387 | -26.6 | 48.0 | 0 | 34.2 |
| Resnet50 | 4.51 | 6.19 | 6.207 | 37.6 | 332.1 | 18412 | 33.7 |
| Inceptionv3 | 6.56 | 8.64 | 8.646 | 31.7 | 391.1 | 14935 | 33.9 |
| Densenet | 3.08 | 5.60 | 5.639 | 83.1 | 547 | 44439 | 33.5 |

Table 2: Performance for inferring a single image with PRIVADO on ten DNN models on a single-core: LeNet [55], VGG19 [56], Wideresnet [57], Resnet110 [58], Resnext29 [59], AlexNet [60], Squeezenet [61], Resnet50 [58], Inceptionv3 [62], and Densenet [63]. Columns 2 – 5 show the performance breakdown, negative value implies that PRIVADO execution time is less than the baseline. Column 6 shows the size of the maximum memory footprint of the enclave observed at runtime. Column 7 shows the NPPS and Column 8 represents the reduction in the TCB w.r.t. to the original Torch library code.

achieve good accuracy on three popular datasets for images: MNIST [45], Cifar-10 [46], and ImageNet [64]. LeNet uses the least number of parameters (62K), and AlexNet uses the highest number (61.1M). The number of layers in these models range from 6 (MLP), all the way to 910 (DenseNet). Lastly, PRIVADO does not reduce the prediction accuracy because (a) the models are trained to convergence in a trusted environment before they are uploaded to the cloud; (b) the PRIVADO transformation does not change the model inference technique or the specific weights and biases. We experimentally confirm that PRIVADO-Generator and PRIVADO-Converter preserve the accuracy of the model; they yields similar accuracies as the original ONNX models on the target datasets.

6.4 Performance

Experimental Setup. We run all our experiments on a machine with 6th Generation Intel(R) Core(TM) i7-6700U processor, 64GB RAM, and 8192 KB L3 cache with Ubuntu Desktop-16.04.3-LTS 64 bits and Linux kernel version 4.15.0-33-generic as the underlying operating system. The BIOS is configured to use the 128 MB for SGX. All our measurements are averaged over 100 iterations and use a single core.

Methodology. For every model, we compute three execution time metrics. The *baseline* time measures standard versions of the model running outside SGX. The *SGX* time measures time to run the model within SGX, but with no obliviousness. The *SGX+CMOV* time measures time to run the model within SGX with added support for obliviousness. During each execution, we record statistics such as the enclave memory size, number of page faults, and input-dependent branches

for each of the models. Table 2 shows the execution time for our benchmark models. PRIVADO incurs overhead between -27.4% and 90.4% . We believe these numbers are acceptable, given the additional security and privacy guarantees. We now describe several interesting results from our evaluation.

SGX-Enclaves improve efficiency for models that fit entirely in SGX memory. Surprisingly, our first finding is that *some models, namely MLP, LeNet, VGG19, Resnet110, and Squeezenet, execute faster with PRIVADO than the baseline* (negative overhead in Table 2). Note that these models are relatively small: they fit entirely within SGX memory (90 MB), and so do not incur page fault overheads. Upon further analysis, we observed that these models run faster with SGX because the SGX SDK provides an efficient implementation of some `libc` functions, such as `malloc`, as compared to standard `libc` used in the baseline execution.

Page faults cause most of the overhead. For models that exceed the 90 MB SGX memory limit, we find that page swapping between the enclave and outside memory (page faults) contributes to most of the performance overhead. To explain this, we use the metric *normalized page faults per second* (NPPS) in Table 2. This metric calculates the number of page faults incurred by the enclave while executing model inference for one image, normalized by the baseline execution time. We observed that the performance overhead is highly correlated with the NPPS metric for all models. Among models that exceed 90MB, AlexNet has the highest NPPS (65245) and the highest overhead (90.4%) while Wideresnet has the lowest NPPS (5080) and lowest overhead (15.6%). These results suggest that for SGX inference, there are *optimization opportunities in model design and implementation that can*

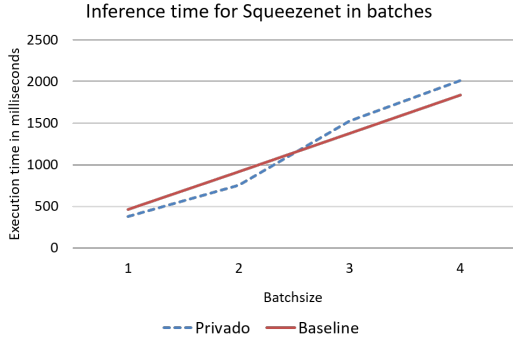


Figure 6: Execution time vs. batch size plot shows that Squeezenet execution time overhead increases with increase in the batch size.

carefully minimize memory size and/or page faults.

Obliviousness is cheap for neural networks. Observe columns *SGX* and *SGX+CMOV* in Table 2. The former captures *SGX* overheads without obliviousness (including decryption of input which costs only about 84 and 313 microseconds for *Cifar-10* and *ImageNet* images respectively), while the latter captures *SGX* overheads after adding obliviousness. We find that *the overhead of adding obliviousness is a mere 1% on average for these ten DNN models*. This observation confirms our claim that neural networks have relatively few input-dependent branches and that the bulk of the computation is access independent, and hence the cost of achieving obliviousness is negligible.

Batching may be counterproductive. We perform an additional experiment to understand the performance overhead for inferring multiple images in a batch. Batching is typically used to improve inference performance. Figure 6 shows the execution time for inferring images in batch sizes of one to four for the Squeezenet model. As shown in Table 2, Squeezenet incurs a negative overhead for a single image as compared to the baseline execution. However, as the batchsize increases, we observe that the execution time for PRIVADO exceeds that of baseline for batchsize of three images and onwards. An increase in the batch size results in larger enclave memory size as the number of activations increases proportionately with batch size. For a batch size of three, Squeezenet memory usage exceeds 90 MB and thus the overhead switches from negative to positive. This result suggests the following rule of thumb: *while performing inference in SGX enclaves, to minimize performance overhead, a smaller batch size that limits memory usage below 90MB is essential.*

Comparison to Previous Work. We compare PRIVADO to previous work by Ohrimenko et al. [7] in terms of performance overhead, expressiveness, and mechanism for obliviousness. We observe that the overhead of Ohrimenko et al. due to *SGX* and encryption ranges from 0.01% to 91% while

that for PRIVADO ranges from -27.4% to 90.4% . Similarly, for the neural network model on *MNIST* dataset, the overhead is around 0.3% due to obliviousness for them which is the same (i.e., 0.3519% with an absolute execution time of 0.000571 sec) for the MLP model with PRIVADO (as shown in Table 2). The overhead due to obliviousness for each model depends on the number of data-dependent layers and the neurons in each layer present in that model. Next, Ohrimenko et al. demonstrate expressiveness by manually modifying 5 distinct machine learning algorithms selected from different categories while PRIVADO focuses specifically on deep learning algorithms and is more expressive for this class of neural networks. PRIVADO automatically converts any given convolutional neural network to run obliviously in the enclave as we modify all the data-dependent layers including pooling layers that are not handled in previous work. At last, although the main technique in PRIVADO to ensure obliviousness is similar in several previous works [7, 23, 24, 47], we leverage our key observation of *assign-or-nothing* patterns in neural networks to design a fully-automated system. Thus, our novel method of applying the oblivious primitive to neural networks combined with PRIVADO-Converter and PRIVADO-Generator is the key contribution of this paper. Moreover, the techniques in PRIVADO are not limited to convolutional neural networks but are suitable for neural networks such as RNNs or LSTMs that exhibit similar data-dependent access patterns.

6.5 Lowering TCB

A naive implementation of PRIVADO would require trusting the entire Torch math and NN libraries, which have approximately 30,000 lines of code. However, PRIVADO-Generator further lowers the TCB using the technique described in Section 5.3. To calculate the reduction in the TCB, we count the number of lines used to generate each of the 10 model binaries. The last column, “TCB Reduction”, in Table 2 shows the percentage reduction in TCB as compared to trusting the entire torch library of 30,000 lines for each model. On average, PRIVADO *results in a 34.7% reduction in TCB for our benchmark models.*

6.6 Obliviousness

PRIVADO-Converter transforms the program to ensure that the execution trace is the same for all inputs. To empirically evaluate this claim, we trace all the instructions executed by the model binary and record all the memory read/write addresses. We build a tracer using a custom PinTool based on a dynamic binary instrumentation tool called Pin [65]. Our tracer logs each instruction before it is executed, along with the memory address accessed (read/write) by the instruction. We run our models before and after using PRIVADO-Converter and log the execution traces for all the inputs in our dataset for checking obliviousness.

| Execution Trace for Airplane | | Execution Trace for Horse | |
|------------------------------|--------------------------------|---------------------------|--------------------------------|
| 0x8915dc | test al, al | 0x8915dc | test al, al |
| <u>0x8915de</u> | <u>jz 0x8915eb</u> | <u>0x8915de</u> | <u>jz 0x8915eb</u> |
| 0x8915e0 | mov rax, qword ptr [rbp-0xf8] | 0x8915eb | mov eax, dword ptr [rbp-0x94] |
| 0x7fffffff118 | ... | 0x7fffffff17c | ... |
| 0x8915e7 | mov eax, dword ptr [rax] | | |
| <u>0xb7939c</u> | ... | | |
| 0x8915e9 | jmp 0x8915f1 | | |
| 0x8915f1 | mov rdx, qword ptr [rbp-0x140] | 0x8915f1 | mov rdx, qword ptr [rbp-0x140] |

(a) Dissimilar execution traces of original code

| Execution Trace for Airplane | | Execution Trace for Horse | |
|------------------------------|-------------------------------|---------------------------|-------------------------------|
| 0x8915d7 | fld st0, dword ptr [rbp-0xac] | 0x8915d7 | fld st0, dword ptr [rbp-0xac] |
| 0x7fffffff174 | ... | 0x7fffffff174 | ... |
| 0x8915dd | fld st0, dword ptr [rbp-0xb0] | 0x8915dd | fld st0, dword ptr [rbp-0xb0] |
| 0x7fffffff170 | ... | 0x7fffffff170 | ... |
| 0x8915e3 | fld st0, dword ptr [rdx] | 0x8915e3 | fld st0, dword ptr [rdx] |
| <u>0xb793ac</u> | ... | <u>0xb793ac</u> | ... |
| 0x8915e5 | fcomi st0, st1 | 0x8915e5 | fcomi st0, st1 |
| 0x8915e7 | fcmove st0, st2 | 0x8915e7 | fcmove st0, st2 |
| 0x8915e9 | fstp dword ptr [rax], st0 | 0x8915e9 | fstp dword ptr [rax], st0 |

(b) Oblivious execution traces with Privado

Figure 7: Execution traces for LeNet on two different inputs (a) before and (b) after applying PRIVADO. The underlined code in (a) shows a conditional jump (jz). The highlighted trace snippet depicts the assign-or-nothing pattern executed for Airplane but not for Horse. The bold and underlined part shows the address which causes a page-fault that the OS can use to distinguish if the branch was executed or not. (b) After PRIVADO transformation, the traces are the same and hence oblivious.

Figure 7 (a) shows two execution trace snippets of LeNet on two different inputs with labels (airplane and horse) before applying PRIVADO-Converter (same as Section 3.2, Figure 3). Observe that the left-hand side trace executes more and different instructions that access different addresses in different modes (read / write) as compared to the right-hand side trace. The adversary can thus distinguish between two inputs by observing such differences in the execution trace. This empirically reinstates that existing library implementations indeed leak input information.

Figure 7 (b) shows the traces after PRIVADO-Converter, they are identical after the transformation. This confirms that PRIVADO-Converter fixed the branches which leaked information in Figure 7 (a). As an empirical evaluation, we collect such execution traces for all the models and their corresponding inputs in the dataset and check the traces. We report that in our experiments, we do not detect any deviation. Hence, we confirm experimentally that we did not miss any branches at least for all the models and inputs in our evaluation.

7 Related Work

In this section, we discuss prior-work on secure neural-network inference using: (a) cryptographic primitives such as homomorphic encryption; and (b) trusted hardware.

7.1 Cryptographic Primitives

CryptoNets [66] was the first to use homomorphic encryption to support neural network inference on encrypted data. Following Cryptonets, several other solutions such as DeepSecure [67], Minionn [68], SecureML [69], ABY3 [70], SecureNN [71], and Gazelle [72], Chameleon [73], Tapas [74] have been proposed for secure neural network inference. These solutions use a combination of cryptographic primitives such as garbled circuits, secret-sharing, and fully homomorphic encryption. Thus, the solutions provide stronger security guarantees as they do not trust any hardware entity. However, these solutions use heavy-weight cryptography and hence incur a significant performance overhead while limiting the ease of use thus making them difficult to adopt in practice.

PRIVADO takes an orthogonal approach to these solutions and uses trusted hardware as the main underlying primitive. Although trusted hardware-based approaches are not a silver bullet and have their share of limitations, we used them because they are more suitable for practical deployments. In comparison to cryptographic primitive-based solutions, PRIVADO provides practical execution performance relying on the availability of trusted hardware and the correct implementation of a secure channel and attestation mechanisms. The choice of using cryptographic primitives or trusted hardware solutions depend on the trade-off that is acceptable between security and performance guarantees.

7.2 Trusted Hardware

Ohrimenko et al. propose a customized oblivious solution for machine learning algorithms using SGX [7]. As discussed in Section 6.4, it does not address the ease-of-use challenge. Recent work Myelin proposes the use of SGX primarily to secure the training process using differential privacy to achieve obliviousness guarantees [75]. PRIVADO, on the other hand, focuses on supporting end-to-end inference-as-a-service for a given trained model. Further, unlike PRIVADO, Myelin is not backward-compatible i.e., it cannot execute inference for models that are *not* trained using the Myelin framework. Our use of the popular Torch framework and ONNX along with automated tools for ensuring obliviousness with ease-of-use differentiates PRIVADO from previous work.

Hunt et al. proposed Chiron, a privacy-preserving machine learning service using the Theano library and SGX [9]. Chiron does not prevent leakage via access patterns which is a serious concern as shown in Section 3. Similarly, MLCapsule, a system for secure but offline deployment of ML as a service on the client-side is susceptible to leakage of sensitive inputs via access patterns [8]. Lastly, Slalom combines SGX and GPU for efficient execution of neural network inference but also does not address access-pattern leakage [10].

Finally, in the light of various side-channel attacks on Intel SGX, several recent works have proposed point-wise solutions to each source of leakage via software and / or hardware modifications [30, 33, 36, 37, 76, 77]. The main goal of these defenses is to be generic-enough to protect a large class of enclave applications at the cost of performance degradation, increased developer-effort, or limited expressiveness of enclave-bound applications. PRIVADO, on the other hand, is tailored for a specific class of enclave applications—deep neural net inference. Our scope and problem setting allow us to strike a balance between various design trade-offs. Several other proposals to thwart side-channel leakage cover a larger attack surface and hence can prevent part of the leakage that we demonstrate via our assign-or-nothing pattern [78]. However, PRIVADO assumes a worst-case adversary and shows that even then it can eliminate such leakage by making the enclave access-pattern oblivious.

8 Conclusion

In this work, we demonstrate the first-of-its-kind concrete attack to highlight the importance of hiding access patterns in DNN inference. Using real-world models executing on a real system, we show that an adversary that observes enclave access patterns can predict encrypted inputs with 97% and 71% accuracy for MNIST and CIFAR10 datasets respectively. To this end, we present PRIVADO—a system which provides secure DNN inference. PRIVADO is input-oblivious, easy to use, requires no developer effort, has low TCB, and has low performance-overheads. We implement PRIVADO on the

Torch framework, apply it to 11 contemporary networks, and demonstrate that PRIVADO is practical and secure.

References

- [1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, 2015.
- [2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, 2012.
- [3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [4] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis.” in *USENIX Security Symposium*, 2016.
- [5] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015.
- [6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *Security and Privacy (SP), 2017 IEEE Symposium on*, 2017.
- [7] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa, “Oblivious multi-party machine learning on trusted processors,” in *25th USENIX Security Symposium (USENIX Security 16)*, 2016.
- [8] L. Hanzlik, Y. Zhang, K. Grosse, A. Salem, M. Augustin, M. Backes, and M. Fritz, “Mlcapsule: Guarded offline deployment of machine learning as a service,” *arXiv preprint arXiv:1808.00590*, 2018.
- [9] 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.
- [10] F. Tramer and D. Boneh, “Slalom: Fast, verifiable and private execution of neural networks in trusted hardware,” *arXiv preprint arXiv:1806.03287*, 2018.
- [11] Asylo, “Introducing asylo: an open-source framework for confidential computing,” <https://asylo.dev/>, 2018.

- [12] M. ACC, “Microsoft azure confidential computing,” <https://azure.microsoft.com/en-us/blog/introducing-azure-confidential-computing/>, 2018.
- [13] S. Shinde, Z. L. Chua, V. Narayanan, and P. Saxena, “Preventing page faults from telling your secrets,” in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 2016, pp. 317–328.
- [14] Y. Xu, W. Cui, and M. Peinado, “Controlled-channel attacks: Deterministic side channels for untrusted operating systems,” in *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, 2015, pp. 640–656.
- [15] S. Lee, M.-W. Shih, P. Gera, T. Kim, H. Kim, and M. Peinado, “Inferring fine-grained control flow inside sgx enclaves with branch shadowing,” *arXiv preprint arXiv:1611.06952*, 2016.
- [16] S. Tople and P. Saxena, “On the trade-offs in oblivious execution techniques,” in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2017.
- [17] A. Baumann, M. Peinado, and G. Hunt, “Shielding applications from an untrusted cloud with haven,” *ACM Transactions on Computer Systems (TOCS)*, vol. 33, no. 3, p. 8, 2015.
- [18] S. Arnautov, B. Trach, F. Gregor, T. Knauth, A. Martin, C. Priebe, J. Lind, D. Muthukumaran, D. O’Keeffe, M. L. Stillwell *et al.*, “Scone: Secure linux containers with intel sgx,” in *12th USENIX Symp. Operating Systems Design and Implementation*, 2016.
- [19] S. Shinde, D. Le Tien, S. Tople, and P. Saxena, “Panoply: Low-tcb linux applications with sgx enclaves,” in *NDSS*, 2017.
- [20] “Software Guard Extensions Programming Reference.” software.intel.com/sites/default/files/329298-001.pdf, Sept 2013.
- [21] V. Costan and S. Devadas, “Intel sgx explained,” Cryptology ePrint Archive, Report 2016/086, 2016.
- [22] C.-C. Tsai, D. E. Porter, and M. Vij, “Graphene-sgx: A practical library os for unmodified applications on sgx,” in *2017 USENIX Annual Technical Conference (USENIX ATC)*, 2017.
- [23] A. Rane, C. Lin, and M. Tiwari, “Raccoon: Closing digital side-channels through obfuscated execution,” in *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, 2015, pp. 431–446.
- [24] S. Sasy, S. Gorbunov, and C. Fletcher, “ZeroTRACE: Oblivious memory primitives from intel sgx,” in *Symposium on Network and Distributed System Security (NDSS)*, 2017.
- [25] torch, “Torch,” <https://github.com/torch>.
- [26] Onnx, “Open neural network exchange format,” <https://onnx.ai/>.
- [27] F. McKeen, I. Alexandrovich, I. Anati, D. Caspi, S. Johnson, R. Leslie-Hurd, and C. Rozas, “Intel® software guard extensions (intel® sgx) support for dynamic memory management inside an enclave,” in *Proceedings of the Hardware and Architectural Support for Security and Privacy 2016*, ser. HASP 2016, 2016.
- [28] T. Knauth, M. Steiner, S. Chakrabarti, L. Lei, C. Xing, and M. Vij, “Integrating remote attestation with transport layer security,” *CoRR*, vol. abs/1801.05863, 2018.
- [29] T. Hunt, Z. Zhu, Y. Xu, S. Peter, and E. Witchel, “Ryoan: a distributed sandbox for untrusted computation on secret data,” in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, 2016, pp. 533–549.
- [30] M.-W. Shih, S. Lee, T. Kim, and M. Peinado, “T-sgx: Eradicating controlled-channel attacks against enclave programs.” in *NDSS*, 2017.
- [31] J. Van Bulck, N. Weichbrodt, R. Kapitza, F. Piessens, and R. Strackx, “Telling your secrets without page faults: Stealthy page table-based attacks on enclaved execution,” in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 1041–1056.
- [32] B. Gras, K. Razavi, H. Bos, and C. Giuffrida, “Translation leak-aside buffer: Defeating cache side-channel protections with {TLB} attacks,” in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 955–972.
- [33] O. Oleksenko, B. Trach, R. Krahn, M. Silberstein, and C. Fetzer, “Varys: Protecting {SGX} enclaves from practical side-channel attacks,” in *2018 {USENIX} Annual Technical Conference ({USENIX}{ATC} 18)*, 2018, pp. 227–240.
- [34] F. Brasser, U. Müller, A. Dmitrienko, K. Kostianen, S. Capkun, and A.-R. Sadeghi, “Software grand exposure: {SGX} cache attacks are practical,” in *11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17)*, 2017.
- [35] J. Götzfried, M. Eckert, S. Schinzel, and T. Müller, “Cache attacks on intel sgx,” in *Proceedings of the 10th European Workshop on Systems Security*. ACM, 2017, p. 2.

- [36] S. Chen, X. Zhang, M. K. Reiter, and Y. Zhang, “Detecting privileged side-channel attacks in shielded execution with déjà vu,” in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 7–18.
- [37] Y. Fu, E. Bauman, R. Quinonez, and Z. Lin, “Sgx-lapd: Thwarting controlled side channel attacks via enclave verifiable page faults,” in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2017, pp. 357–380.
- [38] B. Coppens, I. Verbauwhede, K. De Bosschere, and B. De Sutter, “Practical mitigations for timing-based side-channel attacks on modern x86 processors,” in *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 2009, pp. 45–60.
- [39] J. V. Cleemput, B. Coppens, and B. De Sutter, “Compiler mitigations for time attacks on modern x86 processors,” *ACM Transactions on Architecture and Code Optimization (TACO)*, 2012.
- [40] G. Barthe, T. Rezk, and M. Warnier, “Preventing timing leaks through transactional branching instructions,” *Electronic Notes in Theoretical Computer Science*, 2006.
- [41] “Intel sgx programming reference,” <https://software.intel.com/sites/default/files/managed/48/88/329298-002.pdf>, 2017.
- [42] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.
- [43] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, “Semi-supervised knowledge transfer for deep learning from private training data,” *arXiv preprint arXiv:1610.05755*, 2016.
- [44] M. Nasr, R. Shokri, and A. Houmansadr, “Machine learning with membership privacy using adversarial regularization,” in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 634–646.
- [45] Y. LeCun, “The mnist database of handwritten digits,” <http://yann.lecun.com/exdb/mnist/>.
- [46] cifar, “Cifar-10 dataset,” <https://www.cs.toronto.edu/~kriz/cifar.html>.
- [47] A. Ahmad, K. Kim, M. I. Sarfaraz, and B. Lee, “Obliviate: A data oblivious file system for intel sgx,” in *Symposium on Network and Distributed System Security (NDSS)*, 2017.
- [48] L. N. Huynh, R. K. Balan, and Y. Lee, “Deepsense: A gpu-based deep convolutional neural network framework on commodity mobile devices,” in *Proceedings of the 2016 Workshop on Wearable Systems and Applications*. ACM, 2016, pp. 25–30.
- [49] M. Zhang, S. Rajbhandari, W. Wang, and Y. He, “Deepcpu: serving rnn-based deep learning models 10x faster,” in *2018 USENIX Annual Technical Conference*. USENIX Association, 2018, pp. 951–965.
- [50] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.
- [51] B. Neyshabur, Y. Wu, R. R. Salakhutdinov, and N. Srebro, “Path-normalized optimization of recurrent neural networks with relu activations,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3477–3485.
- [52] T. Reps, S. Horwitz, and M. Sagiv, “Precise interprocedural dataflow analysis via graph reachability,” in *Proceedings of the 22Nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL ’95. ACM, 1995, pp. 49–61.
- [53] A. Brahmakshatriya, P. Kedia, D. P. McKee, D. Garg, A. Lal, A. Rastogi, H. Nameti, A. Panda, and P. Bhatu, “Conflvm: A compiler for enforcing data confidentiality in low-level code,” in *EuroSys*, 2019.
- [54] Onnx, “Onnx model zoo,” <https://github.com/onnx/models>.
- [55] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger *et al.*, “Comparison of learning algorithms for handwritten digit recognition,” in *International conference on artificial neural networks*, vol. 60. Perth, Australia, 1995, pp. 53–60.
- [56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.
- [57] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.
- [58] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [59] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5987–5995.

- [60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [61] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *CVPR*, vol. 1, no. 2, 2017, p. 3.
- [64] imagenet, “Imagenet,” <http://image-net.org/>.
- [65] C.-K. Luk, R. Cohn, R. Muth, H. Patil, A. Klauser, G. Lowney, S. Wallace, V. J. Reddi, and K. Hazelwood, “Pin: Building customized program analysis tools with dynamic instrumentation,” in *Proceedings of the 2005 ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI ’05. New York, NY, USA: ACM, 2005, pp. 190–200.
- [66] 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.
- [67] B. D. Rouhani, M. S. Riazi, and F. Koushanfar, “Deepsecure: Scalable provably-secure deep learning,” in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*. IEEE, 2018.
- [68] J. Liu, M. Juuti, Y. Lu, and N. Asokan, “Oblivious neural network predictions via minionn transformations,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 619–631.
- [69] P. Mohassel and Y. Zhang, “Secureml: A system for scalable privacy-preserving machine learning,” in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 19–38.
- [70] P. Mohassel and P. Rindal, “Aby3: A mixed protocol framework for machine learning.”
- [71] S. Wagh, D. Gupta, and N. Chandran, “Securenn: Efficient and private neural network training.”
- [72] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, “Gazelle: A low latency framework for secure neural network inference,” *arXiv preprint arXiv:1801.05507*, 2018.
- [73] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, “Chameleon: A hybrid secure computation framework for machine learning applications,” in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, 2018, pp. 707–721.
- [74] A. Sanyal, M. J. Kusner, A. Gascon, and V. Kanade, “Tapas: Tricks to accelerate (encrypted) prediction as a service,” *arXiv preprint arXiv:1806.03461*, 2018.
- [75] N. Hynes, R. Cheng, and D. Song, “Efficient deep learning on multi-source private data,” *arXiv preprint arXiv:1807.06689*, 2018.
- [76] V. Costan, I. Lebedev, and S. Devadas, “Sanctum: Minimal hardware extensions for strong software isolation,” in *USENIX Security*, 2016.
- [77] M. Orenbach, Y. Michalevsky, C. Fetzer, and M. Silberstein, “Cosmix: A compiler-based system for secure memory instrumentation and execution in enclaves,” in *USENIX ATC*, 2019.
- [78] M. Werner, T. Unterluggauer, L. Giner, M. Schwarz, D. Gruss, and S. Mangard, “Scattercache: Thwarting cache attacks via cache set randomization,” in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 675–692.