Memory Optimization for Energy-Efficient Differentially Private Deep Learning

Jonathon Edstrom, Hritom Das, *Student Member, IEEE,* Yiwen Xu, Na Gong, *Member, IEEE*

***Abstract*—With the advent of Internet of Things (IoT) technologies and availability of a large amount of data, deep learning has been applied in a variety of artificial intelligence (AI) applications. However, sharing personal data using IoT edge devices carries inherent risks to individual privacy. Meanwhile, the energy and memory resources needed during the inference process becomes a constraint to the resource-limited IoT edge devices. This paper brings memory hardware optimization to meet the tight power budget in IoT edge devices by considering the privacy, accuracy, and power efficiency tradeoff in differentially efficient deep learning systems. Based on a detailed analysis on these characteristics, an Integer Linear Programs (ILP) model is developed to minimize mean square error (MSE), thereby enabling optimal input data memory design. Our simulation results in 45-nm CMOS technology show that the proposed technique can enable near-threshold energy-efficient memory operation for different privacy requirements, with less than 1% degradation in classification accuracy.**

***Index Terms*—Deep learning, embedded memory, power consumption, differential privacy, accuracy, Integer Linear Programs (ILP) model**

I. Introduction

In recent years, deep learning, at the forefront of new developments in artificial intelligence (AI), is transforming many modern applications, from face identification, automatic translation, and computer vision, to self-driving cars, healthcare, and education. For example, deep learning has demonstrated exceptional performance in disease diagnosis of brain disorders and various forms of cancers [1, 2, 3], due to the availability of a large amount of patients' data. Meanwhile, with the advent of wearable technologies and Internet of Things (IoT), there is a rising interest in providing a personalized experience with health recommender systems. For example, smart watches can record cardiac activities [4] and recent medical sensors can replace a finger prick for blood glucose testing [5]. The collected health data can be leveraged through deep learning to provide "personalized" methods of prevention, treatment, and care, thereby aiding persons with disabilities or aging people to address health disparities. As an example, in January 2019, CarePredict, the leading digital health company, launched an AI-powered platform for at-home use by aging seniors. The platform uses deep learning, combined with smart IoT devices, to unobtrusively monitor the daily activities performed by older adults [6].

Such learning-enabled benefit, however, does come with its own cost, such as the associated serious *privacy* concerns. Sharing personal data carries inherent risks to individual privacy. Due to the substantial requirements for computation and storage resources, today's deep learning systems are typically built upon large, centralized data repositories. Based on this centralized-training paradigm, data owners need to upload their private data to the provider and do not have control over how their private data is being used [7, 8, 9].

To protect privacy, one popular technique is differentially private deep learning algorithms [10], which add random noise to the computation so that the output does not significantly depend on any particular training sample (see Fig. 1). When introducing noise, the privacy-guarantee comes at the cost of compromising the accuracy of the models, and this privacy-accuracy trade-off is represented in the differential privacy model through a parameter – privacy budget (ε), which represents the privacy loss in a system. A smaller value of ε indicates a smaller privacy loss (i.e. stronger privacy guarantee) and a larger accuracy degradation of deep learning systems [11, 12]. For a specific application, the value of ε is usually given based on the prior consensus between the users and the deep learning service providers [10, 13].

In addition to privacy, as the deep learning networks grow, the energy and resources needed during the inference process, particularly memories, have become a major constraint to the resource-limited IoT devices [14]. As shown in Fig. 1, during the IoT edge inference process, memory traffic mainly contains two components: (i) weight storage for neural network models and (ii) input data memory to store images, voice, and other sensory signals [15]. In a deep learning accelerator, to store the weights, memory accesses usually consume several orders of magnitude higher energy than computation, making memory performance the bottleneck for processing [16]. For example, in AlexNet, nearly 3000M memory accesses are required, which dominates the power consumption of the entire learning system [16]. In another deep learning Integrated Circuit (IC) -

Manuscript submitted June 1, 2019, revised August 19, 2019, accepted September 2, 2019. This work was supported in part by National Science Foundation under Grant CCF-1855706. (Corresponding author: N. Gong.)

J. Edstrom and H. Das are with the Department of Electrical and Computer Engineering, North Dakota State University, ND 58108 USA (e-mail: jonathon.edstrom@ndsu.edu, hritom.das@ndsu.edu).

Y. Xu is with the Department of Industrial and Manufacturing Engineering, North Dakota State University, ND 58108 USA (e-mail: yiwen.xu@ndsu.edu).

N. Gong is with the Department of Electrical and Computer Engineering, University of South Alabama, AL 36688 USA (e-mail: nagong@southalabama.edu).

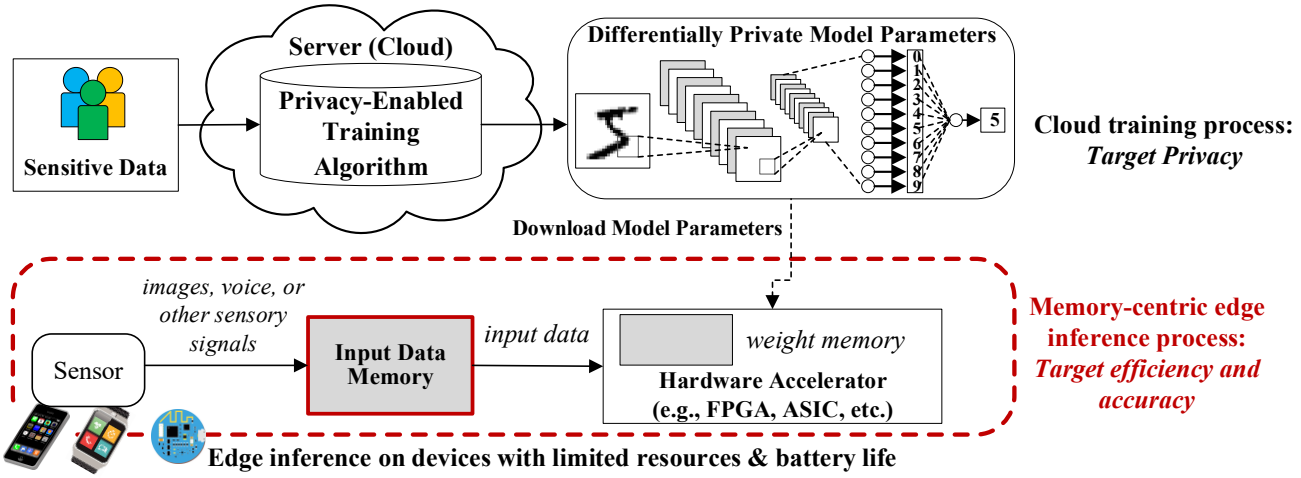


Fig.1. Input data memory optimization for deep learning system with energy-efficiency, privacy, and inference accuracy

DianNao, the static random-access memory (SRAM) occupies 56% of the silicon area and contributes to over 60% power consumption of the entire chip [17]. To reduce the power consumption of weight storage, researchers have developed large embedded SRAM or exploit data reuse to reduce or even eliminate the off-chip memory traffic. ShiDian-Nao [18], for example, is an in-camera CNN accelerator, where the accelerator is placed on the same chip as the image sensor processor, storing all the weights in embedded memory and eliminating off-chip DRAM memory accesses, thereby achieving 60X higher energy-efficiency than DianNao. Additionally, researchers have developed approximate weight memories to optimize the power consumption of the weight storage [19] [20]. As more and more deep learning accelerators adopt embedded memories to store weight values, the input data storage becomes dominant for external memory traffic (e.g., 96.6 mW to read one image compared to 45.3 mW for object detector in an embedded vision system [15]). Consequently, power-efficient embedded input data memory is one of the key design considerations for deep learning systems. Particularly, given the number of IoT devices is predicted to rapidly increase and reaching over 75 billion by 2025 [21], the resulting rapid explosion and scale of collected data brings increasing pressure for input data storage. This paper aims to optimize the power efficiency of the input data storage. We propose a new embedded memory design technique that considers the relationship between efficiency, accuracy, and privacy of differentially private deep learning systems, thereby meeting the increasing storage demands of edge inference. Specifically, this paper makes the following contributions:

- **Input data memory design for differentially private deep learning considering the tradeoff between privacy, accuracy, and power efficiency**: To the best of the authors' knowledge, this is the first work to connect the input data memory to the privacy, accuracy, and power efficiency trade-off in differentially private deep learning systems. We conclude that if the memory hardware can enable the optimal quality of the input data, the accuracy can be optimized for deep learning systems with different privacy levels (Sections III.B and III.C).
- **MSE based mathematical model for optimal memory design**: Based on the general concept of mean square error (MSE), an Integer Linear Programs (ILP) model is developed to enable optimal input data memory design for differentially private deep learning. The developed model can be solved fast by existing solvers (e.g., Gurobi and Cplex), which significantly saves the design time as compared to the traditional ASIC design process (Sections III.D).
- **Novel low-power memory design:** Based on the developed ILP model, a memory design technique is presented that combines the use of other memory techniques (bitcell upsizing and 8T+6T hybrid bitcells). Our analysis shows that significant MSE improvement can be enabled with the optimal design (Section IV-A).
- **Thorough evaluation:** Finally, a comprehensive suite of simulations is performed on the proposed input data memory, and the enriched results include: silicon area constraint, power consumption, data quality, prediction accuracy, and privacy budget. Our evaluation results show that the proposed memory can achieve significant power savings while maintaining near optimal accuracy (details are shown in Section IV).

The organization of the paper is as follows. A review of differentially private deep learning techniques is provided in Section II. Section III analyzes the impact of input data quality and memory in local devices on the system accuracy. Section IV presents the proposed memory hardware design process. The evaluation results are discussed in Section V. Finally, we conclude the paper in Section VI.

II. Learning with Differential Privacy

A. Privacy Preservation in Deep Learning

Privacy research has drawn attention in both industry and research communities. Large industry leaders, including:

Apple, Facebook, and Google, have concluded that these types of threats can be accomplished by invasive analysts even when data has been anonymized [22, 23, 24]. For example, in 2006 AOL released a list of 20 million web search queries which was found to have leaked the identity of a woman [25]. Similarly, Netflix introduced an open competition in 2006 that released a dataset that also leaked private data [26, 27]. One other area with potential privacy issues is biomedical research. For example, in genome wide association studies, the identity and any diseases a person has could be revealed based on results included in research papers [28]. Due to privacy risks such as these, a conscious effort to reduce data leaks has become of great interest, especially for companies using machine learning algorithms on collected big data.

The privacy of deep learning models, such as neural networks, have recently come into question due to weaknesses and attack models that have been previously exploited [29]. Due to high requirements of computation and storage resources, today's deep learning systems are typically built upon large, centralized data repositories. Many cloud providers also give access to computing platforms and learning frameworks for model training, such as Amazon Sagemaker and Google Cloud ML Engine. Based on this centralized-training paradigm, data owners need to upload their private data to the cloud provider and they do not have control over how their private data is being used. For instance, if a deep learning model was trained on the records of patients with a certain disease, learning that an individual's record was part of the training data directly affects their privacy, and it opens a door to potential misuse (e.g., exploitation for the purpose of recruitment, insurance pricing, or granting loans) due to the following three potential privacy threats: (i) it is very easy for a malicious provider to steal the data if the provider has full access to the data [7]; (ii) even without full access to the data, the malicious provider can extract sensitive data from the trained models [8]; and (iii) A malicious remote user can also retrieve information of the training data by carefully querying the trained models [9].

B. Differentially Private Deep Learning and State of the Art

To preserve data privacy, differential privacy [10] is becoming the gold standard to offer both utility to the applications and rigorous privacy guarantees. The formal definition is as follows: a randomized mechanism M is considered to be (ε, δ)-differentially private if, for two adjacent inputs d and d', it holds that $\Pr[M(d) \in S] \leq e^{\varepsilon} \cdot \Pr[M(d') \in S] + \delta$, where S is any subset of the outputs. The privacy cost parameter ε is used to control the tradeoff between the privacy and the accuracy where smaller values of ε provide more privacy. The guarantee of differential privacy is: if an individual's data is used in a differentially private calculation, the probability of any result of the calculation changes by at most a factor of e^{ε} in comparison to if that individual's data is not used in the calculation [30]. The parameter δ is the probability of failure where the given differentially private mechanism may violate an individual's privacy. This δ value explains the possibility of "bad events" that may result in a large loss in privacy. Specifically, when training an (ε, δ) -differentially private neural network, the probability of violating the privacy, δ, is calculated after each step for a given privacy cost, ε.

Recent works have adopted the use of (ε, δ) -differential privacy in order to protect the data of individuals. In [31], the authors presented a technique involving an ensemble of teachers that could train on subsets of a sensitive data. After training, the teachers would further train a student model based on public data that was labeled using the ensemble. The student model is trained based on the noisy voting of the various teachers that were trained using the model so that a stronger privacy guarantee can be enabled by the system. In [32], a method creating generative adversarial networks (GANs) that include differentially private mechanisms to provide privacy guarantees was presented. This technique for training a differentially private GAN only allows the analyst to inspect a model that already guarantees some level of differential privacy. Both the teacher ensemble and differentially private GAN training techniques employ the use of a privacy accountant (i.e. the moments accountant), described in [33], in order to compute a tighter bound on the differential privacy.

In order to ensure differential privacy, perturbation can be introduced at various parts of the workflow, including: input, output, and objective perturbation [34]. Also, different types of noise can be added to the training and test datasets. The moments accountant shows that for the Gaussian (i.e. $\sim N(0, \sigma^2)$) noise mechanism, if the value of standard deviation for this noise mechanism is chosen to be:

$$\sigma = \frac{1}{\varepsilon}\left(2 \log \frac{1.25}{\delta}\right)^{1/2}, \quad (1)$$

then the noise mechanism will satisfy (ε, δ) -differentially privacy for a given sensitivity, S_f. Using this moments accountant technique to compute a tight bound on the privacy allows for each step in the training algorithm to result in (ε, δ)-differential privacy with respect to the lot.

The system we propose uses the moments accountant to train a differentially private ConvNet model on the server (cloud) where sensitive data is used for training. By enabling the moments accountant for training we can guarantee privacy, but at the cost of some accuracy loss. This trained, differentially private model will then be downloaded to the edge computing devices for inference tasks. A diagram of the proposed system design can be seen in Fig. 1. Since inference is taken care of on the local devices, the privacy of the testing data being presented to the devices is not a big concern.

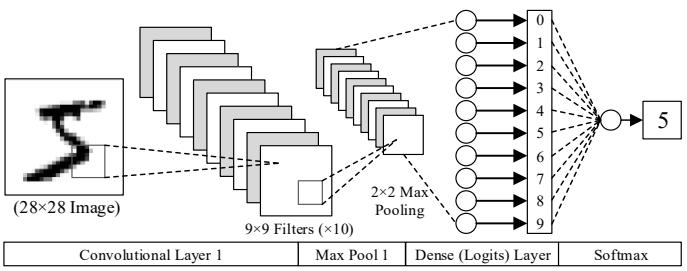


Fig. 2. Differentially private convolutional neural network used in our analysis.

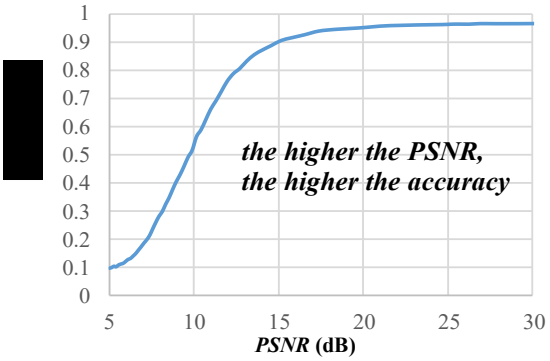


Fig. 3. Influence of dataset quality on test accuracy (using MNIST dataset).

As discussed in Section I, the *energy* and *resources* needed during the inference process has become another constraint to the resource-limited IoT devices. Deep learning models can take up a large portion of an embedded device's memory space and inference tasks. In particular, data movement on these devices can consume the majority of the total power [35]. Software compression techniques for reducing the size of each weight in deep learning models have been introduced, such as the TensorFlow Lite API [36], which allows for 4× reduction in total model size. For hardware improvements, one of the most important issues that has been focused on is the intensive memory access of the embedded IoT devices. Very recently, [37] presented a memory-based noise addition technique for differentially private deep learning systems, illustrating the significance of the embedded memory to edge inference tasks. However, this technique adopted the traditional memory design, which misses out on many optimization opportunities to trade off among privacy, accuracy, and efficiency.

This paper aims to optimize memory design to better support differentially private deep learning algorithms in local devices. To enhance the power efficiency of memories, memory failures are usually introduced due to process variations during the device manufacturing process. We first analyze the impact of memory failures on accuracy and privacy and then conclude the guidelines to optimize the memory for privacy, efficiency, and accuracy in AI applications with different requirements.

III. Impact of Memory Failures in Differentially Private Deep Learning Systems

In our analysis, we define a convolutional neural network model using the TensorFlow framework [38] in order to gain insight on how different types and levels of noise may influence the privacy-accuracy tradeoff. The model involves using an objective perturbation through additive Gaussian noise and uses the moments accountant [33] to compute the privacy cost after each step in the training process. The ConvNet model we tested was based on the architecture described in [37] with a single convolutional layer and can be seen in Fig. 2. The widely used MNIST dataset [39] was used for our initial simulations. MNIST consists of 60,000 training samples and 10,000 test samples, where each sample is a handwritten digit ranging from "0" to "9"; each sample is an image that contains 784 features representing 28×28 pixels.

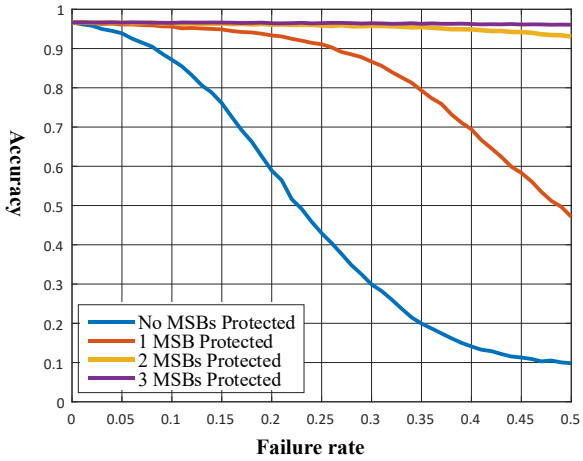


Fig. 4. Impact of memory failure rate on the accuracy of the learning system.

A. Impact of Image Quality on Classification Accuracy

In order to investigate the relationship between the quality of the test dataset and its impact on the test classification accuracy, we inject bit level errors at varying memory failure rates (probabilities) to each image in the test dataset. Since the MNIST dataset consists of images, the well-known peak signal-to-noise ratio (*PSNR*) metric is used to evaluate quality, which is defined in [21] as

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right), \quad (2)$$

where *MSE* is the mean square error between the original images (*Org*) and the degraded images (*Deg*).

Accordingly, by evaluating the *PSNR* values for a wide range of error injected test datasets using MNIST and comparing the test classification accuracy, we identify that the higher the image quality in the test dataset, the higher the output accuracy of the system will be overall. This relationship between *PSNR* and test classification accuracy is illustrated in Fig. 3. Based on this monotonically increasing behavior, if the *PSNR* value of the dataset is improved, the accuracy will be enhanced. Accordingly, during the memory design process, if the memory hardware can enable the optimal quality of the dataset, the accuracy will be improved accordingly. As shown in Fig. 3, as the *PSNR* values of the MNIST dataset are increased from 5dB to 15dB, the accuracy is increased from 10% to 90% while meeting the privacy guarantee for the differentially private deep learning systems. It should be noted that *PSNR* is used in our analysis to evaluate the image quality of MNIST dataset, but considering different types of IoT data, MSE will be a general quality evaluation metric, which will be discussed in Section III-D.

B. Protecting Most Significant Bits (MSBs) of Data

The amount of Gaussian noise that is used during training influences how accurate the inference of the finalized model performs. Therefore, different models with varying amounts of noise (i.e. sigma values) and epsilon values with a set delta value of 10^{-5} have been studied. For sigma, we tested 4 different noise levels, $\sigma \in \mathbf{Z} : 1 \leq \sigma \leq 4$, and for each sigma value we tested 6 separate epsilon values, $\varepsilon \in \mathbf{Z} : 5 \leq \varepsilon \leq 10$.

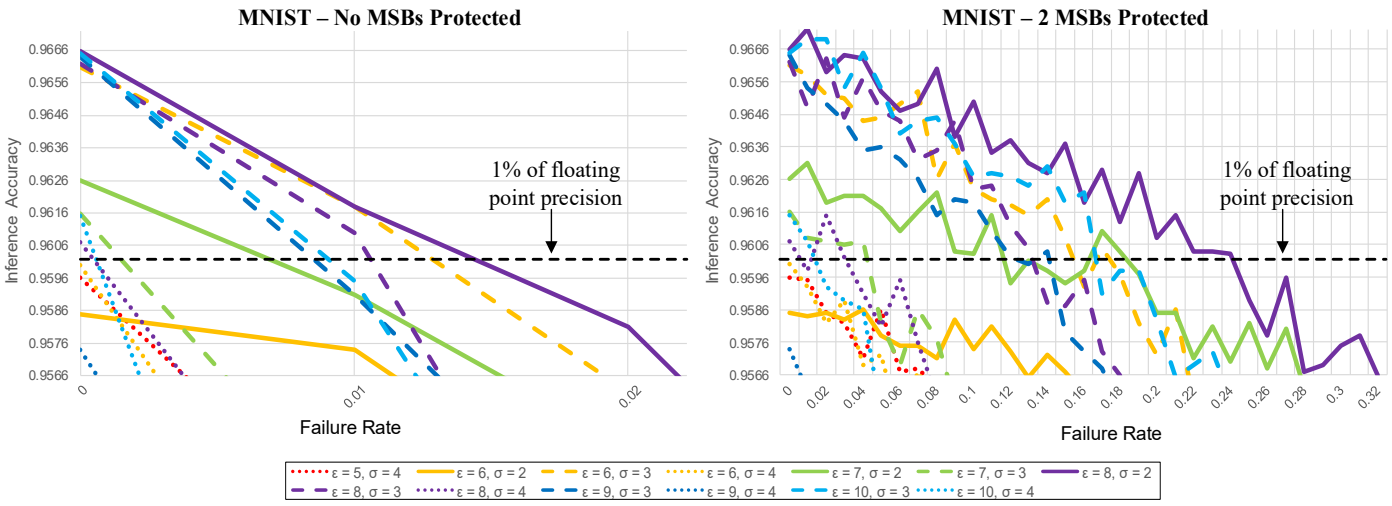


Fig. 5. Impact of memory failure rate on privacy/accuracy tradeoff. (a) without MSB protected and (b) with 2 MSBs protected.

The relevant results for the (σ, ε) pairs we tested are shown in Fig. 5. Our study shows that the best (σ, ε) pairs (i.e. the values of sigma and epsilon that provide the best test classification accuracy) for the MNIST dataset are: $\sigma = 1$, $\varepsilon = 9$ and $\sigma = 2$, $\varepsilon = 8$ as memory failure rates of the dataset are increased. When training using these values for the parameters, the probability of violating the privacy is recalculated after each step in the training process until the end delta value $\delta = 10^{-5}$ to stay within a modest privacy budget [33].

One effective technique for increasing the *PSNR* of the test dataset when errors are present is to protect the most significant bits (MSBs) of the data from memory failures [19, 20]. To study the impact of the memory failures, we further investigate the individual cases of protecting 1, 2, or 3 MSBs and compare against the case without protecting any bits to see the influence of the MSBs on the test classification accuracy. Fig. 4 displays the test classification accuracy of $\sigma = 2$, $\varepsilon = 8$ differentially private ConvNet with the varying amount of MSBs protected. The protection of 2 or 3 bits has a significant influence on boosting the accuracy of the system to acceptable levels.

C. *Impact of Memory Failure on Privacy/Accuracy Trade-off*

Additionally, the impact of the memory failure on the privacy/accuracy trade-off is studied in differentially private deep learning systems. It can be seen from Fig. 5 (a) that, the parameter ε represents the general trade-off between privacy level and accuracy of the differentially private deep learning system. A larger value can potentially enable higher accuracy. Additionally, for this specific Gaussian (i.e. $\sim N(0, \sigma^2)$) noise addition mechanism, the value of σ also directly indicates the trade-off between privacy and accuracy. As shown in Fig. 5 (a), in general, as σ (i.e. the amount of noise) increases, the accuracy decreases.

When comparing Fig. 5 (a) and (b), it can be observed that for an optimal input data memory with MSBs protected the accuracy/privacy tradeoff can be significantly improved. For example, in the case where $\sigma = 2$ and $\varepsilon = 8$, if the memory failure rate is 0.23, without protection, the accuracy will be much less than any acceptable amount (i.e. within 1% of the error free system). By introducing the protection to 2 MSBs, at the same failure rate, the accuracy will be increased to >96%, which is within 1% of the fault free differentially private model. In the following section, based on the design guidelines, a low power memory will be designed to minimize power consumption while keeping an acceptable accuracy for the differentially private deep learning systems.

D. *Integer Linear Programs (ILP) Model based Memory Design*

Based on the above analysis, we propose an input data memory design technique to improve the prediction accuracy of differentially private deep learning systems. To optimize the dataset quality, the design problem becomes an energy-accuracy-cost tradeoff design problem. We apply the model for hybrid SRAM without bitcell integration cost (i.e., Model 2) in [40] to handle this problem. In the following we provide an independent brief introduction to the mathematical model.

Assume the data points $y_1, y_2, \cdots, y_n$ are stored in a memory, and each data point needs s memory bitcells to store. Then, the mean square error (MSE) of these data points is defined by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(D)} - y_i^{(O)}\right)^2,$$

where $y_i^{(D)}$ and $y_i^{(O)}$ are the degradated and original data values, respectively. The degradations are caused by hardware memory failures. The expected MSE can be calculated by

$$E(MSE) = \frac{1}{ns}\sum_{i=1}^{n}\sum_{k=1}^{s} 4^k q_{ik},$$

where q_{ik} is the given failure probability of the k^{th} bitcell of the i^{th} data [40]. Note that the general concept of MSE is widely used in data analytics and statistics, not only in image or video pixels.

Suppose we have r_1 and r_2 design options for 6T and 8T SRAM, respectively. Let $r = r_1 + r_2$ be the total design option. In addition, define binary decision variables

$$x_{ikl} = \begin{cases} 1, & \text{if option } l \text{ is chosen for the } ik^{th} \text{ bitcell} \\ 0, & \text{otherwise} \end{cases}$$
$$(i = 1, \cdots, n; k = 1, \cdots, s; l = 1, \cdots, r)$$

Then the following Integer Linear Programs (ILP) model can be formulized to enable an optimal input data memory using 6T sizing techniques and 8T+6T hybrid design

$$\min_{x} \sum_{i=1}^{n} \sum_{k=1}^{s} \sum_{l=1}^{r} 4^k q_{ikl} x_{ikl} \quad (3)$$

s.t. $$\sum_{l=1}^{r} x_{ikl} \geq 1, \ i = 1, \cdots, n; k = 0, \cdots, s \quad (4)$$

$$\sum_{i=1}^{n} \sum_{k=0}^{s} \sum_{l=1}^{r} s_{ikl} x_{ikl} \leq s_{total} \quad (5)$$

$$x_{ikl} \in \{0,1\}, \ i = 1, \cdots, n; k = 1, \cdots, s; l = 1, \cdots, r \quad (6)$$

The objective function (3) is to minimize the expected MSE of the whole data set. Constraint (4) is to guarantee that one can choose exactly one design option for each bit cell. Note that since this is a minimization problem, (4) is equivalent to $\sum_{l=1}^{r} x_{ikl} = 1$. The total area constraint (5) assures that the total area of the design cannot exceed the given limit s_{total}, where s_{ikl} is a known parameter indicating the area cost of the ik^{th} bitcell if it is selected to adopt the l^{th} design option. Finally, constraint (6) indicates that x_{ikl} is a binary decision variable.

The following section will present the memory design and evaluate results in a 45nm CMOS technology based on this optimization model. It should be noted that the developed ILP model can be used for optimal memory design in different technologies.

IV. Embedded Memory Design for Deep Learning

To evaluate the effectiveness of the proposed memory design technique, 0.4V and 0.5V are used in our analysis based on a 45nm CMOS technology to enable the maximum energy efficiency at near-threshold voltage [41, 42]. The deep learning system was set up using a single convolutional layer, a learning rate of 0.05, a batch size of 600, and an L2-norm gradient bound of 4.0 for norm clipping. The total epochs for any given privacy level are calculated during training and are based on the privacy parameters ε and δ, and the noise parameter σ. For example, with large target ε (i.e. less privacy) and/or large σ (i.e. more noise), the network model can be trained for more epochs without violating the chosen privacy level.

A. Optimized Memory Design

Traditional low-power memories often utilize bitcell sizing or more than 6T bitcells to reduce memory failures induced by process variations, thereby achieving power savings at low voltages. This is because, at low voltages, memory failures are mainly caused by the intra-die variations in process parameters (e.g., variations in channel length, channel width, oxide thickness, threshold voltage, line-edge roughness, and random dopant fluctuations [RDF]) and the inter-die variations (i.e. different process corners including "typical NMOS and typical PMOS", "fast NMOS and slow PMOS", "slow NMOS and fast PMOS", "slow NMOS and slow PMOS", and "fast NMOS and fast PMOS") [19, 20, 43]. Among the different sources of intra-die variations, RDF-induced threshold voltage (V_{th}) variations are the most significant in causing memory failures [43], which can be expressed by

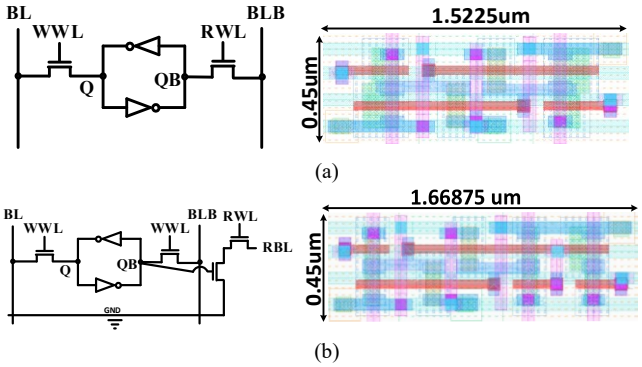


Fig. 6. Different memory designs: (a) 6T SRAM schematic and minimum-sized layout design in 45 nm technology (C61) and (b) 8T SRAM schematic and minimum-sized layout design (C81) in the same 45 nm technology.

TABLE I. MEMORY FAILURE RATE

memory bitcells	height (μm)	width (μm)	area (μm^2)	area ratio s_k	*failure rate* @0.4V	*failure rate* @0.5V
6T: C61	0.45	1.523	0.685	1	0.5897	0.3436
6T: C62	0.45	1.563	0.703	1.026	0.5341	0.3074
6T: C63	0.45	1.603	0.721	1.053	0.4803	0.2771
6T: C64	0.45	1.643	0.739	1.079	0.4342	0.2521
8T: C81	0.45	1.669	0.751	1.096	0.0121	0.00082
8T: C82	0.45	1.700	0.765	1.117	0.0043	0.00009
8T: C83	0.45	1.740	0.783	1.143	0.002	0.00002

$$\sigma V_{th} = \sigma V_{th0} \sqrt{\frac{W_{min} \ L_{min}}{W \ L}}, \quad (7)$$

where σV_{th0} is the standard deviation of V_{th}, and W, L represent the width and length of the transistor. σV_{th} for an NMOS and PMOS transistor with W equal to the minimum L_{EFF} in the 45nm predictive technology is 46.9mV and 41.8mV, respectively. According to (7), σV_{th} is inversely proportional to $\sqrt{WL}$, indicating that the deviation of V_{th} is reduced as W and L increase. Therefore, upsizing bitcells can reduce memory failures at low voltages due to the reduced intra-die threshold voltage (V_{th}) variations.

In addition to upsizing bitcells, more than 6T bitcells can also mitigate process variation caused memory failures. Fig. 6 shows the 6T bitcell and 8T bitcell width using 45 nm CMOS technology. As shown, 6T bitcells can achieve better area-cost performance while 8T can effectively reduce memory failures due to its decoupled read and write paths using two extra transistors. However, 8T bitcells causes about 9.6% area overhead compared to 6T.

Memory failure rates are also strongly dependent on inter-die variations. Under inter-die variations, the dominant failures of 6T and 8T occur in read operations at "fs" (fast NMOS and slow PMOS) and in write operations at "sf" (slow NMOS and fast PMOS) process corners, respectively [20]. In our analysis, 10,000 Monte-Carlo simulations are performed with local intra-die threshold voltage variations (RDF effects) at the worst process corners for 6T and 8T bitcells. The failure rates are listed in Table I. As shown, there are $r = r_1 + r_2 = 4 + 3 = 7$ total options (including 4 upsized 6T options and 3 upsized 8T options). As expected, upsizing bitcells and 8T options can both result in a lower failure rate with a larger bitcell area. Also, as

TABLE II. RESULTS AND COMPARISONS

S_{total}	Optimal Design @ 0.5V									Traditional Scenario		*MSE* Improvement
	$MSE_{opt.}$	S_7	S_6	S_5	S_4	S_3	S_2	S_1	S_0	$MSE_{Trd.}$	$Des._{optn.}$	
8.0	12034.27	C61	C61	C61	C61	C61	C61	C61	C61	12034.27	C61	0.00%
8.1	3003.22	C81	C61	C61	C61	C61	C61	C61	C61	12034.27	C61	75.04%
8.3	200.77	C81	C81	C81	C61	C61	C61	C61	C61	10337.20	C62	98.06%
8.5	28.56	C81	C81	C81	C81	C81	C61	C61	C61	8997.34	C63	99.68%
8.7	**2.91**	**C83**	**C83**	**C82**	**C81**	**C81**	**C81**	**C61**	**C61**	**7944.11**	**C64**	**99.96%**
8.9	0.80	C83	C83	C82	C81	C81	C81	C81	C81	18.01	C81	95.56%
9.1	0.43	C83	C83	C83	C83	C83	C83	C82	C82	1.94	C82	77.84%
S_{total}	Optimal Design @ 0.4V									Traditional Scenario		*MSE* Improvement
	$MSE_{opt.}$	S_7	S_6	S_5	S_4	S_3	S_2	S_1	S_0	$MSE_{Trd.}$	$Des._{optn.}$	
8	26207.11	C61	C61	C61	C61	C61	C61	C61	C61	26207.11	C61	0.00%
8.1	6883.95	C81	C61	C61	C61	C61	C61	C61	C61	26207.11	C61	73.73%
8.3	735.63	C81	C81	C81	C61	C61	C61	C61	C61	22596.87	C62	96.74%
8.5	150.27	C83	C83	C82	C81	C61	C61	C61	C61	19329.43	C63	99.22%
8.7	**56.60**	**C83**	**C83**	**C82**	**C81**	**C81**	**C81**	**C61**	**C61**	**16710.36**	**C64**	**99.66%**
8.9	45.46	C83	C83	C83	C83	C82	C81	C81	C61	270.28	C81	83.18%
9.1	43.80	C83	C83	C83	C83	C83	C83	C82	C82	94.57	C82	53.69%

TABLE III. POWER CONSUMPTION OF OPTIMIZED MEMORY AT 45NM CMOS TECHNOLOGY @ 0.5V

S_{total}	Proposed optimal design		Traditional design			$P_{reduction}$ @ 0.4v (opt.) vs. 1v (Trd.)	$P_{reduction}$ @ 0.5v (opt.) vs. 1v (Trd.)
	$P_{opt.}$ (W) @ 0.4V	$P_{opt.}$ (W) @ 0.5V	$P_{Trd.}$ (W) @ 0.4V	$P_{Trd.}$ (W) @ 0.5V	$P_{Trd.}$ (W) @ 1.0V		
8	1.30E-06	2.07E-06	1.30E-06	2.07E-06	9.28E-06	86.03%	77.69%
8.1	1.41E-06	2.53E-06	1.30E-06	2.07E-06	9.28E-06	84.85%	72.74%
8.3	1.63E-06	3.01E-06	1.34E-06	2.15E-06	1.00E-05	83.74%	69.90%
8.5	1.74E-06	3.50E-06	1.38E-06	2.29E-06	1.16E-05	85.02%	69.83%
8.7	**1.96E-06**	**3.55E-06**	**1.42E-06**	**2.42E-06**	**1.41E-05**	**86.11%**	**74.82%**
8.9	2.07E-06	4.09E-06	2.18E-06	4.22E-06	1.02E-04	97.97%	95.99%
9.1	2.18E-06	3.85E-06	2.18E-06	3.87E-06	1.02E-04	97.86%	96.23%

$P_{opt.}$: power consumption of the proposed memory; $P_{Trd.}$: power consumption of traditional memory design; $P_{reduction}$: power reduction

the supply voltage decreases from 0.5V to 0.4V, the memory failure rate of the same bitcell increases accordingly.

Solving (3)-(6) for a variety of s_{total} values in the range [8.0, 9.1] using Gurobi solver (version 7.0.2) at both 0.4V and 0.5V, the results are listed in Table II. In the traditional design, all bitcells select the same option as discussed in [40]. It can be seen that in most design cases significant MSE improvement can be enabled with the optimal design, including over 99% MSE improvement for both 0.4V and 0.5V if the total area constraint is 8.5 or 8.7. Another interesting observation is that for two different voltages, the optimization solutions under the same total area constraint have the same tendency: when s_{total} is small (e.g., < 8.3), the most cost-efficient bitcell (C61) is usually selected to meet the area constraint. In the extreme case, with s_{total} = 8.0, all bitcells are C61, which is the only possible solution under such a strict area constraint. As s_{total} increases beyond 8.5, a larger number of different 8T bitcells are selected to optimize the quality. It should be noted that as s_{total} = 8.5 or 8.9, the optimal solutions for 0.4V are different from the ones for 0.5V. This is because, for different memory bitcells, the relationship between memory failure and voltage may not be linear [20].

B. Power Efficiency

We have also evaluated the power efficiency of the optimized memory design, as displayed in Table III. All possible memory operations were considered for the total power estimation, including: read (i.e. read zero and read one), write (i.e. write zero to zero, zero to one, one to zero, and one to one), and hold (i.e. leakage power while holding zero and leakage power while holding one). As shown in Table III, operating at 0.4V enables significant power savings as compared to the traditional supply voltage (1V). As the total area constraint S_{total} increases, the power consumption increases due to more 8T bitcells being included in the optimized design solution. If 8.7 is the target area constraint, then 74.82% and 86.11% power savings can be enabled at 0.5V and 0.4V compared to 1V, respectively.

C. Input Data Quality and Accuracy

We further evaluate the input data quality and prediction accuracy using the optimized memory. The results are listed in Table IV. The MNIST dataset [39], which was used as the original dataset for training the CNN model, displays almost no accuracy loss (0.01%) as compared to the fault free test samples. Additionally, the Fashion [44] and Kuzushiji-MNIST (KMNIST) [45] datasets are introduced to evaluate the efficiency of the proposed technique. The Fashion and KMNIST datasets are comprised of 28×28 grayscale images of 70,000 fashion product and Japanese characters respectively, with each dataset containing samples from 10 categories. In both datasets the training set has 60,000 images and the test set has 10,000 images. Both Fashion and KMNIST datasets serve as drop-in replacements for the MNIST dataset, as they share the same image sizes and number of classes. The complexity of the Fashion dataset is considered to be moderately more complex to classify than the MNIST dataset while KMNIST is significantly more complex. This level of dataset complexity is

reflected in the classification accuracy results. When training a CNN model with the same architecture on the Fashion dataset, the proposed memory yields a negligible accuracy loss when voltage scaling to 0.5V (0.03%) or 0.4V (0.33%). When training a CNN model with the same architecture using the KMNIST dataset, a dataset that is significantly more difficult to classify, the proposed memory still yields negligible loss in classification accuracy when voltage scaling to 0.5V (0.15%) or 0.4V (0.59%).

The results in Table IV are based on the specific privacy level where the maximum accuracy is enabled for the MNIST dataset (i.e. $\sigma = 2, \varepsilon = 8$). With the same privacy level, using the proposed memory design, the accuracy almost remains the same for the Fashion and KMNIST datasets while the supply voltage is reduced from 1V to 0.4V. The Fashion and KMNIST datasets display higher accuracies for lower levels of ε, but still maintain high accuracy for varying levels of noise.

D. Accuracy at Different Privacy Levels

To evaluate the impact of privacy levels on the effectiveness of the proposed memory technique, varying σ and ε values are included in CNN model simulations. It shows that the privacy level has a noticeable impact on the inference accuracy of the differentially private deep learning systems. The MNIST,

TABLE IV. INPUT DATA QUALITY AND ACCURACY

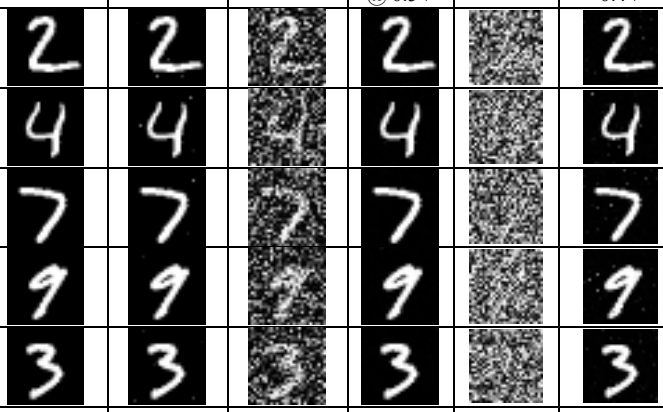
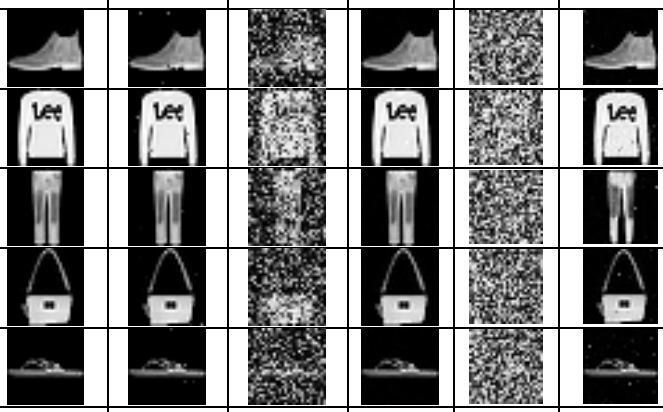
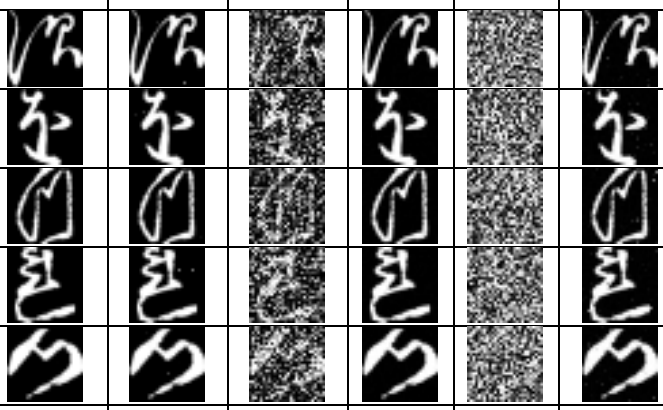
	No Error	1V Trd.	0.5V Trd.	This Work @ 0.5V	0.4V Trd.	This Work @ 0.4V
MNIST Dataset [39]						
Test Accuracy ($\sigma = 2, \varepsilon = 8$)	96.7%	96.67%	42.3%	96.69%	12.22%	96.6%
Fashion Dataset [44]						
Test Accuracy ($\sigma = 2, \varepsilon = 8$)	87.1%	87.06%	31.75%	87.07%	11.59%	86.77%
KMNIST Dataset [45]						
Test Accuracy ($\sigma = 2, \varepsilon = 8$)	80.16%	79.87%	36.84%	80.01%	14.99%	79.57%

TABLE V. IMPACT OF PRIVACY LEVEL ON TEST ACCURACY

Dataset	Privacy Parameters	Privacy /Noise Level	1V Trd.	0.5V Trd.	This Work @ 0.5V	0.4V Trd.	This Work @ 0.4V
MNIST	$\sigma = 4, \varepsilon = 5$	High	95.89%	35.36%	95.91%	13.66%	95.74%
	$\sigma = 2, \varepsilon = 10$	Low	96.52%	48.49%	96.39%	14.15%	96.34%
Fashion	$\sigma = 4, \varepsilon = 5$	High	86.33%	27.53%	86.4%	11.25%	86.1%
	$\sigma = 2, \varepsilon = 10$	Low	87.54%	20.14%	87.64%	10.33%	87.16%
KMNIST	$\sigma = 4, \varepsilon = 5$	High	81.38%	25.14%	81.46%	11.11%	81.29%
	$\sigma = 2, \varepsilon = 10$	Low	83.01%	36.13%	82.98%	13.89%	82.69%

Fashion, and KMNIST datasets were used to determine the impact of the privacy level on the inference accuracy. In general, the higher the privacy level is, the lower the test accuracy becomes. This relationship can be seen in Table V, which includes both high and low levels of privacy for comparison of test accuracy calculations. As displayed in Table V, the proposed memory design at both 0.4V and 0.5V performs similarly to the 1V traditional design and is capable of achieving inference accuracy within 1% of the fault free model at both low and high privacy levels. Therefore, the proposed memory can be a preferable solution for implementing power-efficient differently private deep learning systems.

V. Conclusion

This paper analyzed the power efficiency, accuracy, and privacy characteristics of differentially private deep learning systems and presented an input data memory design with upsized devices and 8T+6T hybrid bitcells, thereby achieving power efficiency and accuracy optimization at different privacy levels. It concluded that the memory design, which achieves the optimal quality of the input data, can provide the highest prediction accuracy with different privacy levels. To enable the presented design technique, a mean squared error (MSE) based Integer Linear Programs (ILP) model was developed for optimal memory design with different silicon area constraints in differentially private deep learning systems, which significantly saved design time as compared with traditional time-consuming and laborious ASIC design processes. Simulation results demonstrate significant reduction in power consumption under different silicon area design constraints, with less than 1% degradation in classification accuracy for different privacy levels. Future investigations would include extension of the proposed optimal memory design to deal with activation private data storage in partitioned deep learning systems (e.g., [13]).

References

[1] J. Kent, "Google Deep Learning Tool 99% Accurate at Breast Cancer Detection," Health IT Analytics, 22 October 2018. [Online]. Available: https://healthitanalytics.com/news/google-deep-learning-tool-99-accurate-at-breast-cancer-detection. [Accessed 22 May 2019].

[2] B. Siwicki, "Johns Hopkins researchers use deep learning to combat pancreatic cancer," Healthcare IT News, 16 August 2018. [Online]. Available: https://www.healthcareitnews.com/news/johns-hopkins-researchers-use-deep-learning-combat-pancreatic-cancer. [Accessed 22 May 2019].

[3] L. Hu, D. Bell, S. Antani, Z. Xue, K. Yu, M. P. Horning, N. Gachuhi, B. Wilson, M. S. Jaiswal, B. Befano, L. R. Long, R. Herrero, M. H. Einstein, R. D. Burk, M. Demarco, J. C. Gage, A. C. Rodriguez, N. Wentzensen and M. Schiffman, "An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening," *Journal of the National Cancer Institute,* 2019.

[4] "Apple Watch Series 4," Apple, 2019. [Online]. Available: https://www.apple.com/apple-watch-series-4/health/. [Accessed 22 May 2019].

[5] "FDA approves first blood sugar monitor without finger prick," CBS News, 29 September 2017. [Online]. Available: https://www.cbsnews.com/news/fda-approves-first-blood-sugar-monitor-without-finger-prick/. [Accessed 22 May 2019].

[6] "CarePredict Launches AI-Powered Platform for Seniors Aging at Home, at CES 2019," CarePredict, 8 January 2019. [Online]. Available: https://www.carepredict.com/news/carepredict-launches-ai-powered-platform-for-seniors-aging-at-home-at-ces-2019/. [Accessed 22 May 2019].

[7] C. Song, T. Ristenpart and V. Shmatikov, "Machine Learning Models that Remember Too Much," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, 2017.

[8] 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,* vol. 10, no. 3, pp. 137-150, 2015.

[9] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *IEEE Symposium on Security and Privacy*, San Jose, 2017.

[10] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science,* vol. 9, no. 3-4, pp. 211-407, 2014.

[11] J. Lee and C. Clifton, "How Much Is Enough? Choosing epsilon for Differential Privacy," in *14th Information Security Conference (ISC'11)*, Xi'an, 2011.

[12] A. C. Valdez and M. Ziefle, "The Users' Perspective on the Privacy-Utility Trade-Offs in Health Recommender Systems," *International Journal of Human-Computer Sutides,* vol. 121, pp. 108-121, 2018.

[13] J. Wang, J. Zhang, W. Bao, X. Zhu, C. B and P. Yu, "Not Just Privacy: Improving Performance of Private Deep Learning in Mobile Cloud," in *KDD*, 2018.

[14] J. Gage, "Machine Learning models on the edge: mobile and IoT," Medium, 20 June 2018. [Online]. Available: https://heartbeat.fritz.ai/machine-learning-models-on-the-edge-mobile-and-iot-8a5384a370ba. [Accessed 22 May 2019].

[15] L. Guo, D. Zhao, J. Zhou, S. Kimura and S. Goto, "Lossy Compression for Embedded Computer Vision Systems," *IEEE Access,* vol. 6, pp. 39385-39397, 2018.

[16] V. Sze, Y.-H. Chen, T.-J. Yang and J. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE,* vol. 105, no. 12, pp. 2295-2329, 2017.

[17] T. Chen, D. Z, N. Sun, J. Wang, C. Wu, Y. Chen and O. Temam, "DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning," in *ASPLOS '14*, 2014.

[18] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun and O. Temam, "Dadiannao: A machine-learning supercomputer," in *47th Annual IEEE/ACM International Symposium*, 2014.

[19] S. Gopalakrishnan, P. Wijesinghe, S. S. Sarwar, A. Jaiswal and K. Roy, "Significance driven hybrid 8T-6T SRAM for energy-efficient synaptic storage in artificial neural networks," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, 2016.

[20] J. Edstrom, Y. Gong, D. Chen, J. Wang and N. Gong, "Data-Driven Intelligent Efficient Synaptic Storage for Deep Learning," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 64, no. 12, pp. 1412-1416, 2017.

[21] "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)," [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. [Accessed August 2019].

[22] A. Chin and A. Klinefelter, "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study," *North Carolina Law Review,* vol. 90, no. 5, 2012.

[23] J. Tang, A. Korolova, X. Bai, X. Wang and X. Wang, "Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12," *ArXiv,* 2017.

[24] Ú. Erlingsson, V. Pihur and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Scottsdale, 2014.

[25] M. Barbaro and T. Zeller, "A Face Is Exposed for AOL Searcher No. 4417749," The New York Times, New York, 2006.

[26] D. Jackson, "The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever," Thrillist.com, 2017.

[27] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *IEEE Symposium on Security and Privacy*, Oakland, 2008.

[28] R. Wang, Y. F. Li, X. Wang, H. Tang and X. Zhou, "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, Chicago, 2009.

[29] W. M. Holt, "Security and Privacy Weaknesses of Neural Networks," Provo, 2017.

[30] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce and A. Roth, "Differential Privacy: An Economic Method for Choosing Epsilon," in *IEEE 27th Computer Security Foundations Symposium*, Vienna, 2014.

[31] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow and K. Talwar, "Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data," in *5th International Conference on Learning Representations*, Toulon, 2017.

[32] X. Zhang, S. Ji and T. Wang, "Differentially Private Releasing via Deep Generative Model," *ArXiv,* 2018.

[33] 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*, Vienna, 2016.

[34] A. D. Sarwate and K. Chaudhuri, "Signal Processing and Machine Learning with Differential Privacy," *IEEE Signal Processing Magazine,* pp. 86-94, September 2013.

[35] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman and Z. Zhang, "Hardware for Machine Learning Challenges and Opportunities," in *IEEE Custom Integrated Circuits Conference*, Austin, 2017.

[36] Google, "TensorFlow Lite," Google, 2018. [Online]. Available: https://www.tensorflow.org/lite/. [Accessed 3 December 2018].

[37] L. Yang and B. Murmann, "Approximate SRAM for Energy-Efficient, Privacy-Preserving Convolutional Neural Networks," in *IEEE Computer Society Annual Symposium on VLSI*, Bochum, 2017.

[38] Google, "TensorFlow ™," Google, [Online]. Available: https://www.tensorflow.org/. [Accessed 11 11 2018].

[39] Y. LeCun, C. Cortes and C. J. Burges, "THE MNIST DATABASE of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/. [Accessed 20 March 2019].

[40] Y. Xu, H. Das, Y. Gong and N. Gong, "On Mathematical Models of Optimal Video Memory Design," *IEEE Trans. on Circuits and Systems for Video Technology,* early access.

[41] N. Gong, J. Edstrom, D. Chen and J. Wang, "Data-Pattern Enabled Self-Recovery Multimedia Storage System for Near-Threshold Computing," in *IEEE International Conference on Computer Design (ICCD'16)*, Scottsdale, 2016.

[42] A. Ferrerón, D. Suárez-Gracia, J. Alastruey-Benedé, T. Monreal-Arnal and P. Ibáñez, "Concertina: Squeezing in Cache Content to Operate at Near-Threshold Voltage," *IEEE Trans. On Computers,* vol. 65, no. 3, pp. 755-769, 2016.

[43] S. Mukhopadhyay, H. Mahmoodi and K. Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.,,* vol. 24, no. 12, p. 1859–1880, 2005.

[44] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," 28 August 2017. [Online]. Available: https://arxiv.org/abs/1708.07747. [Accessed 3 April 2019].

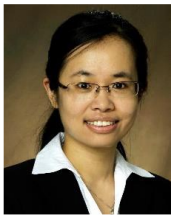
[45] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto and D. Ha, "Deep Learning for Classical Japanese Literature," 3 December 2018. [Online]. Available: http://www.arxiv.org/pdf/1812.0118.pdf. [Accessed 10 August 2019].

Jonathon Edstrom received the B.S. degree in computer engineering, the M.S. degree in Electrical and Computer Engineering from North Dakota State University (NDSU), Fargo, North Dakota, in 2015 and 2017, respectively. Currently, he is pursuing his Ph.D. degree in Electrical and Computer Engineering at North Dakota State University, Fargo, ND. His research focuses on embedded systems implementation of low-power mobile video and vision technology.

Hritom Das received the B.S. degree in electrical and electronic engineering from American International University-Bangladesh, Dhaka, Bangladesh, in 2012 and the M.S. degree in electronic engineering from Kyungpook National University, Daegu, South Korea, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering at North Dakota State University, Fargo, ND, USA.

From 2016 to 2017, he was a faculty member (lecturer) with Uttara University, Uttara, Bangladesh. His research interest includes the low power circuit design, testing, machine learning implementation in traditional electronics.

Yiwen Xu received the Ph.D. degree in systems and industrial engineering from the University of Arizona. He is currently an Assistant Professor in the Department of Industrial and Manufacturing Engineering at North Dakota State University, Fargo, ND. His research interests include applied operations research (especially probabilistic network optimization and applied integer programming) and reliability engineering.



Na Gong (M'13) received the B.E. degree in electrical engineering, the M.E. degree in microelectronics from Hebei University, Hebei, China, and the Ph.D. degree in computer science and engineering from the State University of New York, Buffalo, in 2004, 2007, and 2013, respectively.

Currently, Dr. Gong is an Associate Professor with the Department of Electrical and Computer Engineering at University of South Alabama, Mobile, AL, USA. Her research interests include power-efficient computing circuits and systems, memory optimization, and neuromorphic hardware.