Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Overlapped Data Processing Scheme for Accelerating Inference Procedure in Machine Learning

JINSEO CHOI[1] AND DONGHYUN KANG[2]
[1]Changwon National University, 20 Changwondaehak-ro Uichang-gu Changwon-si, Gyeongsangnam-do (51140), South Korea
[2]Changwon National University, 20 Changwondaehak-ro Uichang-gu Changwon-si, Gyeongsangnam-do (51140), South Korea

Corresponding author: Donghyun Kang (e-mail: donghyun@gs.cwnu.ac.kr).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1I1A3047006)

ABSTRACT

With the enhancement of machine learning (ML) technologies, many researchers and developers have challenged how to solve the traditional problem using ML technologies with a rich set of hardware resources. Unfortunately, the resources are still wasted even though there are many efforts to study the internal processing of ML so as to maximize the provided resources as much as possible. In this paper, we first observe possible optimization opportunities by studying the data flows of the conventional inference procedure in detail. Then, we propose `ol.data`, the first software-based data processing scheme that aims at (1) overlapping training steps with validation steps inside one epoch or two adjacent epochs and (2) performing validation steps in parallel; it helps to significantly enhance not only computation time but also resource utilization. By using `ol.data`, we implement CNN model and compare it with the traditional approaches, Numpy (i.e., baseline) and tf.data, under three differnt datasets. The results show that `ol.data` improves inference performance by up to 41.8% while using the highest CPU and GPU utilization.

INDEX TERMS Machine learning, TensorFlow, CPU/GPU utilization, Overlapping, Multiple threads

I. INTRODUCTION

FOR several years, machine learning (ML) has achieved great success in diverse computing environments [1]–[3]. Emerging hardware technologies (e.g., Ampere [4], Apple M1 [5], TPU [6], etc.) and open-source libraries for supporting ML (e.g., TensorFlow [2], [7], PyTorch [8], Caffe [9], Keras [10], etc.) had been helpful to accelerate the successful adoption of ML theories and models. For example, the inference procedure now becomes a standard feature in ranking, image classification, or recommendation systems [11], [12]. The procedure commonly runs on a machine that consists of CPUs and lots of GPUs; CPUs transform data into *tensors* and GPUs perform *training* and *validation* steps to make a prediction with high accuracy. Unfortunately, such an inference procedure often wastes rich hardware resources (e.g., CPU cycles and memory) as all steps should be fully serialized before starting the next epoch; an epoch means iteration granularity where the entire dataset is trained and validated once.

To solve the aforementioned issue, there were many interesting approaches [13]–[16]. Some researchers focused on the same issue and they proposed tf.data that processes some of the inference steps (i.e. data load and preprocessing) in a parallel way [13]. Unfortunately, tf.data still misses opportunities in exploiting parallelism inside one or multiple epochs. Meanwhile, recent research offloads time-consuming parts of the inference procedure into hardware logic (i.e., FPGA) to sufficiently utilize the hardware resources [1], [3], [17]–[22]. For example, Ren et al. focused on GPU-CPU hybrid training technology that offloads data to reduce the amount of data movement to/from GPU [20]. Kwon et al. introduced the prototype of the multi-FPGA acceleration for large-scale RNNs [21]. However, we think that the offloading approaches are difficult to be gained wide adoption in various environments because they require not only new hardware but also the modification of the corresponding codes.

In this paper, we propose a new data processing scheme, called overlapped data (`ol.data`), to efficiently address the resource waste issue. The key insight of `ol.data` is that there is a sequence of the inference procedure (e.g., data-

load, preprocessing, training, and validation) and the traditional procedure inevitably wastes the hardware resources by running each step with a single thread. To keep high CPU and GPU utilization, (1) `ol.data` logically isolates and overlaps the training and validation steps by using *model copy* operation inside adjacent epochs and (2) it allows multiple threads to perform the validation steps concurrently. For evaluation, we implemented CNN model for image classification by extending tf.data based on TensorFlow framework which is widely used to implement ML models and extensively evaluated `ol.data` with three popular datasets, such as MNIST, fashion MNIST, and Extended MNIST. The major contributions of this paper are the following.

- We first describe data flows of the conventional inference procedure in detail and carefully observe how much CPU and GPU resources are utilized in the entire inference model.
- We propose `ol.data`, which reduces the idle time to start the next step by overlapping the CPU and GPU executions and saves the computing time consumed by the validation step using multiple threads.
- Finally, we show comprehensive evaluation results to confirm the effectiveness of `ol.data`. For comparison, we additionally implemented CNN models with Numpy and tf.data on TensorFlow framework, respectively. Experimental results show that `ol.data` reduces the elapsed time consumed by the inference procedure by up to 41.8% compared with tf.data.

The rest of this paper is organized as follows. Section II introduces how the inference procedure behaves to correctly make a prediction and describe and describes our motivation from new observation results. Section III explains the key techniques of `ol.data`. Section IV shows our evaluation and comparison with existing data processing approaches. Finally, Section V introduces related work and Section VI concludes this paper.

II. BACKGROUND AND MOTIVATION

In this section, we describe how to transform data into tensors and run the training and validation operations in detail. We introduce our new observations from the monitoring of CPU and GPU utilization during the inference procedure and motivation of this work.

A. DATA FLOWS IN ML INFERENCE PROCEDURE

In ML domains, raw input data is transformed through a designated procedure and transferred from CPU to GPU to predict a correct answer.

Fig. 1 presents the data processing flow of the entire inference procedure inside one epoch. In general, the inference procedure can be classified into five steps: data-load, preprocessing for training, training, preprocessing for validation, and validation. (1) In data-load step, raw data is loaded as an argument of input through either networks or file systems; this step only runs at the initial time to

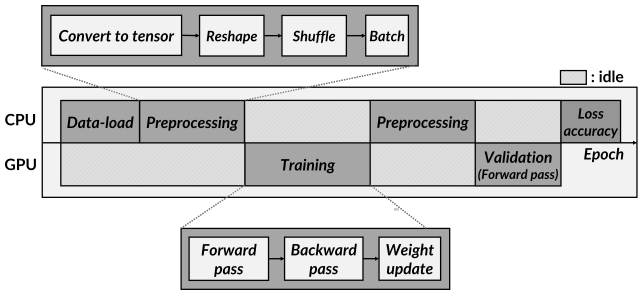


FIGURE 1: The data flow of the conventional inference procedure.

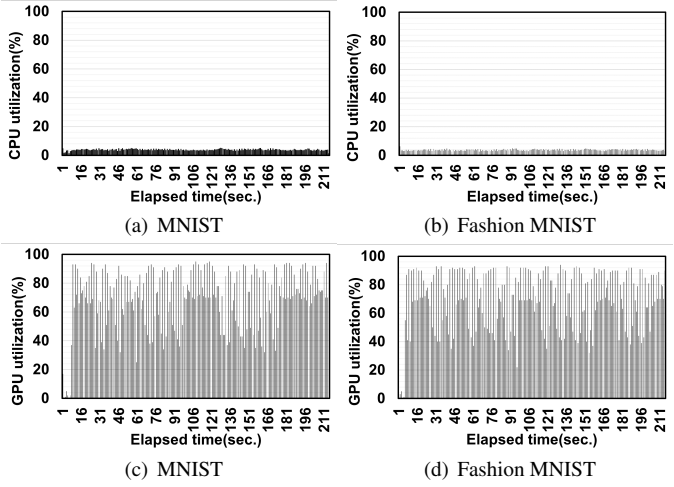


FIGURE 2: CPU and GPU utilization for CNN model (100 epochs)

start. (2) Next, Some loaded data from the previous step are transformed into tensors for training; about 80% of the whole data is commonly used in this step. The tensors are reshaped and shuffled to avoid erroneous inference and to achieve higher accuracy. Then, a set of tensors is delivered as training arguments for ML models (e.g., CNN [23], RNN [24], and DNN [25]) by going through a batch operation. (3) To find correct answers from the corresponding ML model with tensors, The training step runs three operations including forward pass, backward pass, and weight update. (4) Then, to perform preprocessing for the next validation step, the remaining data about 20% of the loaded data is transformed in the same way as before. (5) Finally, the validation step makes acceptable predictions based on the weight learned in the training step. After finishing this step, the loss and accuracy values are calculated to confirm that the model reaches an acceptable loss and accuracy.

Note that, the steps are conducted in a serialized fashion because one thread is responsible for each step; the next step is conducted after finishing the prior step. As shown in Fig. 1, CPU and GPU resources are not used even though they are idle because each step waits for the completion of the previous step. As a result, a rich amount of hardware resources are unintentionally wasted.

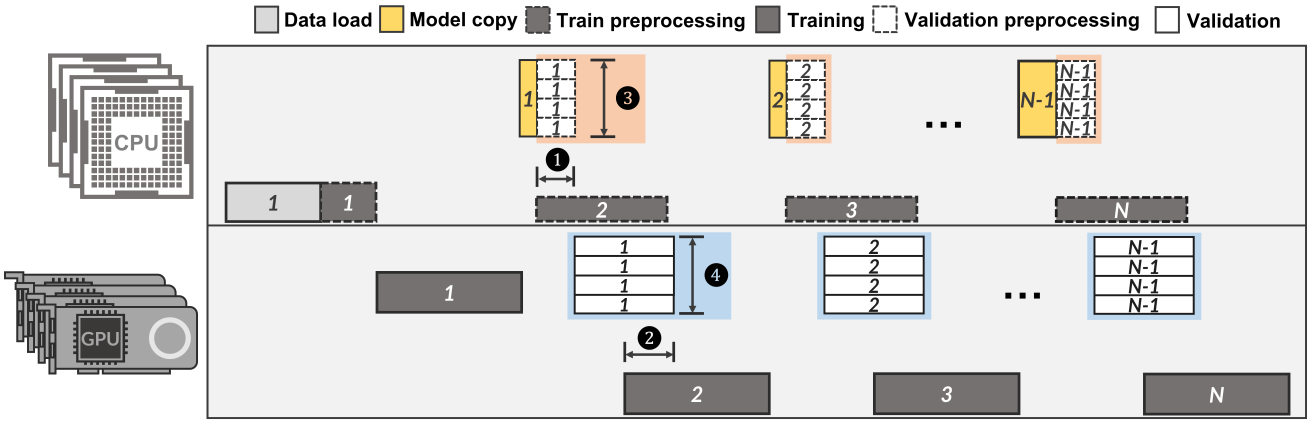


FIGURE 3: The process of ol.data. The number inside each rectangle means the corresponding epoch time.

B. CPU AND GPU UTILIZATION

In order to confirm our intuition as mentioned before, we conducted experiments to study CPU and GPU utilization. In this evaluation, we implemented image classification model based on TensorFlow with Numpy and followed the default guidelines to build the model with MNIST and fashion MNIST datasets; the datasets are widely used in image classification models [26], [27].

Fig. 2 shows evaluation results of utilization measured while running the model on a machine composed of CPUs and GPUs (we will describe the evaluation setup in detail later). Unfortunately, as shown in Fig. 2(a) and Fig. 2(b), the model utilizes about 3.7% of CPU resources in both datasets. This observation is very interesting in that the steps running on CPU do not fully utilize the available resources, even though there is no interference with other tasks. To understand why CPU has inferior utilization, we conducted a study and found that (1) the utilization collapse at the start time comes from the I/O overhead in the data-load step and (2) most parts of inferior utilization are derived from that an order of the four operations in the preprocessing step is executed based on a single thread. In addition, some operations require a lot of time and efforts For example, the shuffle operation first reads an array of NumPy to get the index of each tensor and randomly exchanges an order of the whole tensors. Fig. 2 also shows a pattern in which the utilization often drops because the preprocessing step is triggered after finishing the GPU operations.

Meanwhile, utilization of GPU shows better than that of CPU but reveals dramatic fluctuations (see Fig. 2(c) and Fig. 2(d)). The reason behind this is that the training and validation step should wait for the completion of the preprocessing inside one epoch and this wait is repeated every epoch. In summary, the serialized step in the traditional inference procedure can waste a rich set of CPU and GPU resources. This limitation motivates the need for intelligent isolation and parallelism scheme.

III. DESIGN

In this section, we propose a new data processing scheme to guarantee high CPU and GPU utilization while running the inference procedure, which we call `ol.data`. The design goal of `ol.data` is to accelerate the overall performance of ML model while efficiently isolating tasks running on CPU or GPU. Fig. 3 shows the data processing flow of `ol.data` in detail.

A. DATA-LEVEL ISOLATION

As mentioned in Section II-A, steps belonging to the same inference model are serialized using a single thread even though they run on physically independent cores. The key reason is that a few tasks on CPU or GPU share the data structures for managing model parameters (i.e., *weight values*). In general, a task in the validation step employs the parameters which have been finally updated by the previous training task; since the model parameters are frequently updated during the execution of the training step, the value of the parameters in the middle of the execution is incomplete to use the validation task.

To efficiently disable this data sharing, `ol.data` has an additional operation between training and validation steps, called *model copy* operation. The operation provides a simple and powerful solution in that it can isolate and overlap the training task and validation task with low overheads. As shown in Fig. 3, `ol.data` performs three steps including data-load, preprocessing for training, and training step, in the same way as processing the traditional ones. Then, `ol.data` executes the model copy operation in which the whole data structures updated by the previous training task are copied for enabling the *data-level isolation*. Finally, `ol.data` additionally creates new tasks (i.e., threads) and runs them so as to separate the flow of the rest steps on the current epoch from the main task. The created tasks simultaneously perform validation based on the copied model parameters after finishing the corresponding preprocessing step (we will describe how `ol.data` handle them in detail later). Meanwhile, the main task starts the next epoch by triggering the preprocessing

step for training regardless of the process in the validation tasks. Such a concurrent execution never interferes with the flow and calculation of the separated validation tasks running on CPU and GPU because `ol.data` efficiently takes the data-level isolation. As a result, `ol.data` can provide opportunities to improve the overall performance in that two or more steps can be overlapped. In other words, `ol.data` can concurrently run two preprocessing steps; one is for the training of the next epoch and the other is for the validation of the current epoch (❶). The execution of the validation step can be overlapped with that of the training step as much as possible (❷). We believe that such a parallelism improves the overall performance with high CPU and GPU utilization.

Meanwhile, if the training step for the next epoch completes before the validation step in the current epoch, the model copy operation is postponed until finishing the validation step; it helps to reduce the management overhead of the shared data structures.

B. ENHANCED VALIDATION STEP

As the aforementioned before, a long completion time for the validation step might be harmful in that the model copy operation is delayed. To reduce the inevitable wait, we enhanced the validation step to further optimize `ol.data`. The key idea of the enhancement is to (1) split data used in the validation step and (2) efficiently schedule tasks derived from the related steps using multiple threads. Therefore, the enhanced validation step has extra operations for making acceptable predictions.

Fig. 3 also shows how `ol.data` enhances the validation step along with the corresponding preprocessing step. (1) After a model copy operation, `ol.data` creates concurrent tasks (i.e., threads) by up to the predefined number of threads; we limit the number of overlapped threads to avoid interference overheads across tasks. (2) `ol.data` equally splits the whole data used by the preprocessing step into smaller chunks to allocate them to the created tasks one by one. Data belonging to each chunk is preprocessed by the corresponding task in parallel (❸). Therefore, `ol.data` can reduce the time consumed by the preprocessing step compared with the conventional step in a batch. (3) The tasks concurrently perform the validation step using the data from the preprocessed step to calculate predictions. (❹). Note that in the enhanced validation step, each task only performs to make predictions for the allocated data in the corresponding chunk based on the copied data structures. As a result, the execution of the enhanced validation step can have an opportunity to be overlapped on GPU as much as possible. Unfortunately, `ol.data` can be forced to stay in this step for a while so as to finish all tasks because the completion time of each task is different. (4) Finally, `ol.data` reports the loss and accuracy of this validation step by computing the average of all tasks; it is required because the calculated loss and accuracy derived from each task are not visible to other tasks. We believe this operation incurs negligible collection overhead.

In summary, the idea of `ol.data` is simple and straightforward, but it has opportunities to optimize the overall performance with high CPU and GPU utilization because it efficiently provides the data-level isolation and enhanced validation step.

IV. EVALUATION

In this section, we shows our experimental results in detail. Specially, our evaluation aims to answer the following questions:

- How effective is `ol.data`'s parallelism in CPU and GPU utilization?
- Does high utilization derived from benefits of `ol.data` help to improve performance under the various workloads (i.e., datasets) of ML?
- How does `ol.data` perform when the dataset size is large?
- Does `ol.data`'s parallelism negatively affect the loss or accuracy of ML model?

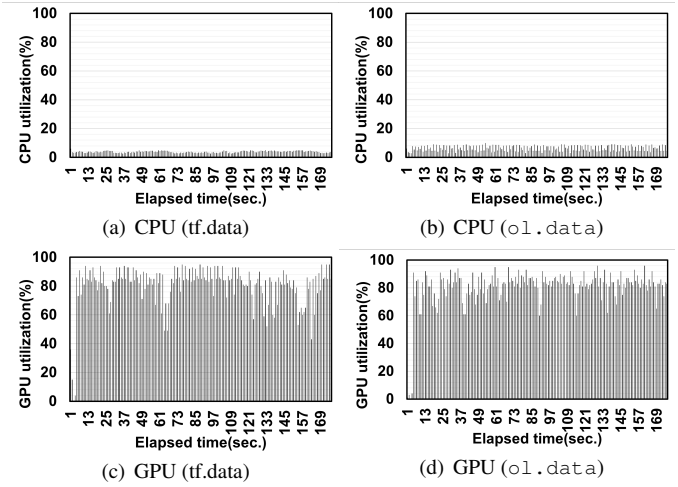
A. EXPERIMENTAL SETUP

For evaluation, we implemented `ol.data` based on tf.data which is one of the state-of-the-art approaches realizing the inference model. We compared it with two traditional approaches, such as Numpy and tf.data; Numpy is well-known as a default implementation (i.e., baseline) and tf.data allows to overlap some steps with multiple threads. Our experiments have been performed by implementing a well-known CNN inference model. For a comprehensive evaluation, we repeated the same experiment by using three different datasets: MNIST (127MB) [26], Fashion MNIST (155MB) [27], and Extended MNIST (1.6GB) [28]. All experiments were run 64-bit Linux on a machine with Intel i9 CPU and NVIDIA GTX 1650 GPU. Table 1 lists the details of a machine and software environment.

TABLE 1: Hardware and software environment.

	Description
CPU	Intel® Core™ i9-12900KF
GPU	NVIDIA GeForce™ GTX 1650
Memory	32GB
Storage	Intel® 660P SSD (512GB)
OS	Ubuntu 20.04.1 LTS(64-bit)
TensorFlow	version 2.3.0
Keras	version 2.8.0

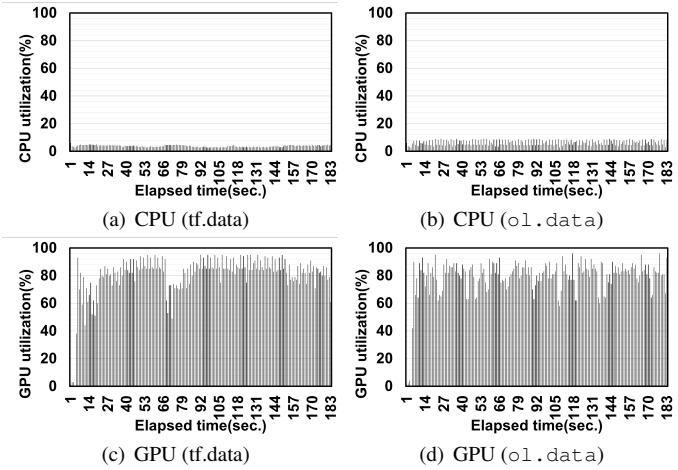
For a comprehensive analysis, inference was run for 100 epochs where one epoch is a full pass over the dataset. As mentioned before, tf.data and `ol.data` can optimize some training and validation steps by configuring the number of threads. Thus, We set the number of threads as 24; it is the same as the number of physical CPUs. Finally, we used `sysstat` [29] and `gpustat` [30] to measure resource utilization on CPU and GPU while running training and validation steps.



(a) CPU (tf.data) (b) CPU (ol.data)

(c) GPU (tf.data) (d) GPU (ol.data)

FIGURE 4: Comparison of CPU utilization on MNIST dataset.



(a) CPU (tf.data) (b) CPU (ol.data)

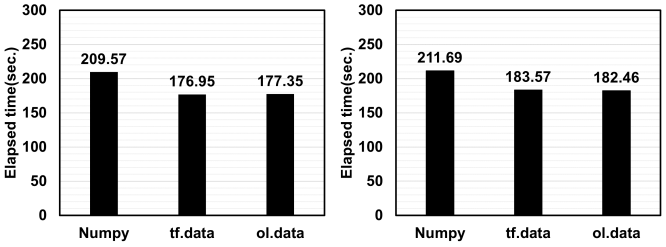
(c) GPU (tf.data) (d) GPU (ol.data)

FIGURE 5: Comparison of GPU utilization on Fashion MINST dataset.

B. DEFAULT MNIST AND FASHION MNIST DATASETS

We first conducted the same experiment mentioned in Section II with `ol.data` and tf.data to confirm the effect of multiple threads. In this experiment, we employ widely used default datasets (i.e., MNIST and Fashion MNIST) to perform CNN model.

Fig. 4 and Fig. 5 illustrate how much `ol.data` and tf.data efficiently can utilize CPU and GPU resources on MNIST and Fashion MINST datasets, respectively. As we expected, they show higher CPU and GPU utilization compared with Numpy where most steps are serialized with a single thread (see Fig. 2). Especially, `ol.data` clearly presents the highest utilization in all kinds of datasets. To clearly confirm the effectiveness of `ol.data`, we measured average CPU and GPU utilization during the training and validation steps. As a result, compared with Numpy on MNIST, `ol.data` improves the overall CPU and GPU utilization by 71% and 16.2%, respectively. The reason behind such improvement is that `ol.data` enables the *enhanced validation step* where



(a) MNIST (b) Fashion MNIST

FIGURE 6: Overall performance comparison on two different datasets.

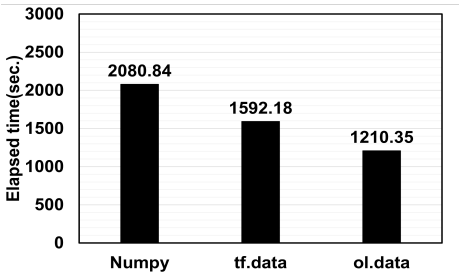


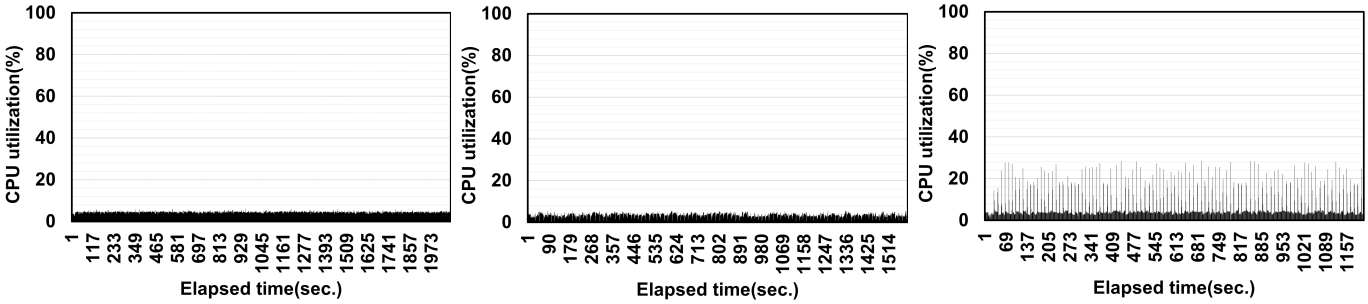
FIGURE 7: Overall performance comparison on Extended MNIST.

tensors can be simultaneously inserted and produced when performing the operations associated with each layer in inference model. In addition, `ol.data` depicts higher utilization of CPU than that of tf.data; it improves by up to 64.9% on MNIST and 68.9% on Fashion MNIST, respectively. The good CPU utilization of `ol.data` mainly comes from the *data-level isolation* where training and validation step no longer waits for each other to finish. Unlike CPU, average GPU utilization of `ol.data` is similar to that of tf.data. To understand the reason, we performed breakdown analysis and found that the overlapped region on GPU is very short because of small size of the dataset.

Next, we validate the overall performance to show whether high CPU and GPU utilization of `ol.data` can lead to performance gain. Fig. 6 shows the elapsed time of Numpy, tf.data, and `ol.data`. Fig. 6 clearly indicates the poor performance of Numpy compared with `ol.data`. Surprisingly, `ol.data` shows comparable performance compared to tf.data even though it took the highest CPU utilization on both datasets. The main reason of the results is that `ol.data` unfortunately reveals the potential inefficiency of the extra operation for supporting parallelism (i.e., model copy and waiting time to calculate of loss and accuracy). However, we believe that these overheads would be negligible as the size of the dataset increases; it increases the region for overlapping of `ol.data` because the validation steps spend more time to complete.

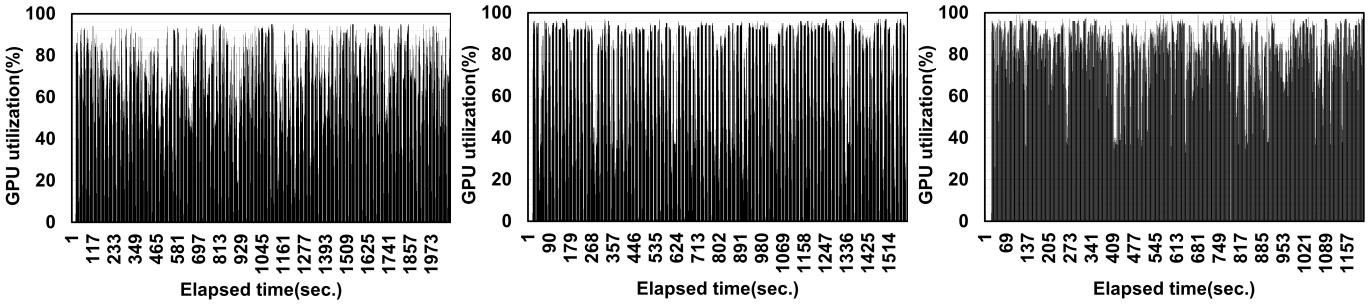
C. EXTENDED MNIST DATASET

To confirm the efficiency of `ol.data`, we additionally conducted the same experiment and analysis on Extended



(a) Numpy (b) tf.data (c) `ol.data`

FIGURE 8: Comparison of CPU utilization on Extended MNIST.



(a) Numpy (b) tf.data (c) `ol.data`

FIGURE 9: Comparison of GPU utilization on Extended MNIST.

MNIST dataset [28]; the dataset size of the Extended MNIST is 12.9 times larger than that of the default MNIST.

Fig. 7 presents the performance results of CNN model under the three different approaches, including Numpy, tf.data, and `ol.data`. As we expected, `ol.data` shows the speedup of CNN model compared with traditional approaches; it achieves performance improvement by up to 41.8% and 24% compared to Numpy and tf.data, respectively. As mentioned before, these results mean that `ol.data` can naturally isolate between training and validation steps with the large overlapped region and the overhead derived from the extra operations is negligible. To understand such a performance gain in detail, we conducted a sensitivity analysis by measuring CPU and GPU utilization on the fly.

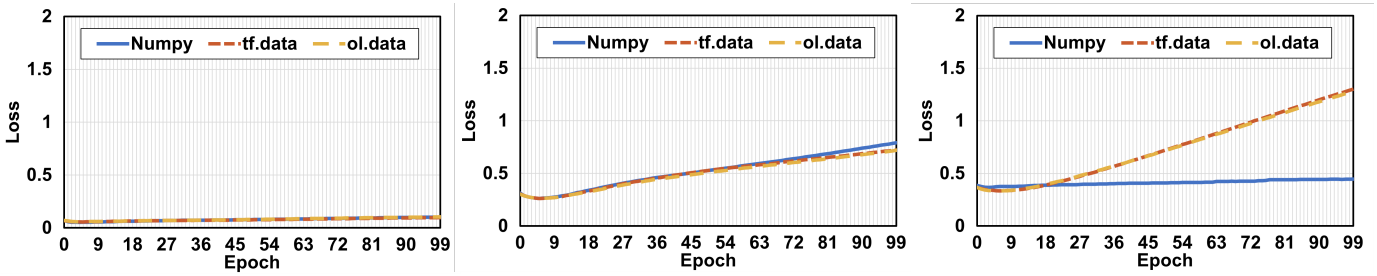
Fig. 8 and Fig. 9 show our empirical results. One interesting result is that `ol.data` utilizes highly CPU resources by up to 28.46% (see Fig. 8(c)). To clearly quantify the impact of high CPU utilization, we calculated average CPU utilization. As a result, we found that CPU utilization of `ol.data` outperforms that of tf.data by 75.7% on average. To analyze such a significant gap, we focused on the gain of `ol.data` in a large dataset and we found the following strengths:

- The large size of the dataset increases the amount of data read and transformed via preprocessing steps of each epoch, thus, the steps should spend more time to be completed on CPU.
- The long processing time leads to more opportunities for computation in parallel because `ol.data` allows to overlap a preprocessing step for validation with a step for inference (see ❶ in Fig. 3).
- In addition, `ol.data` can utilize more resources in that its preprocessing step for validation runs in parallel on CPU (see ❸ in Fig. 3).

On the other hand, the traditional approaches show a similar trend of the results in Section IV-B even though the size of the dataset increases. This is because Numpy and tf.data perform most steps of their counterparts in Fig. 8(c) in a serialized way.

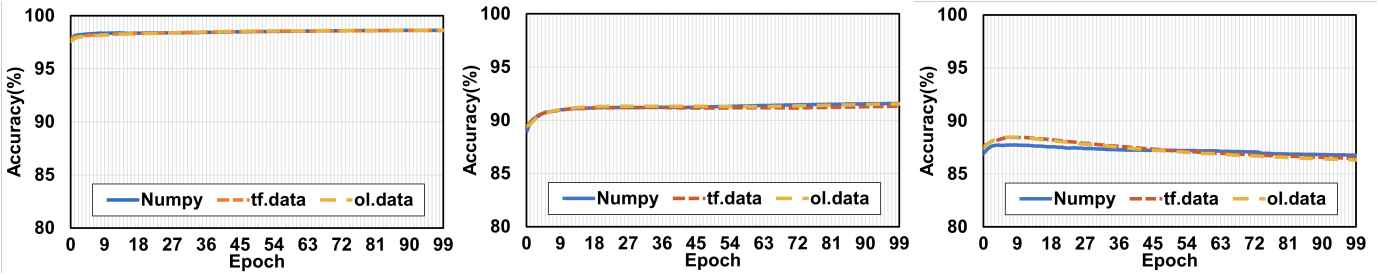
Now, we explain how `ol.data` works to maximize GPU utilization. Fig. 9 presents GPU utilization on Extended MNIST. As shown in Fig. 9(c), `ol.data` often plots a region of high utilization on GPU and it is much more thick and intense compared to that of other figures; average utilization of `ol.data` is 38.7% and 30.2% higher than Numpy and tf.data, respectively. Especially, this result is significantly meaningful in that the gap between `ol.data` and the traditional approaches is larger than that of the result in Section IV-B. We believe that the noteworthy improvement comes from the same strengths as mentioned before (see ❷ and ❹ in Fig. 3).

In summary, `ol.data` can have several benefits: (1) maximizing CPU and GPU utilization by processing crucial steps of inference model in parallel , (2) reducing the amount of time waiting for completion of the prior step.



(a) MNIST (b) Fashion MNIST (c) Extended MNIST

FIGURE 10: Loss pattern for the inference model (100 epochs) on three different datasets. A lower value is better in the model.



(a) MNIST (b) Fashion MNIST (c) Extended MNIST

FIGURE 11: Accuracy pattern for the inference model (100 epochs) on three different datasets. A higher value is better in the model.

D. LOSS AND ACCURACY

Unlike the traditional approaches, `ol.data` performs validation steps with multiple threads. Thus, It needs to merge model loss and accuracy values of each validation task at the end of an epoch; we generate the model loss and accuracy by calculating the average loss and accuracy of all tasks. Note that the merge operation does not cause the huge overhead as described before. But, this behavior in `ol.data` might negatively affect the loss or accuracy of the inference model. To confirm this, we additionally measured the loss and accuracy of Numpy and tf.data.

Fig. 10 and Fig. 11 present model loss and accuracy pattern during 100 epochs. As shown in Fig. 10 and Fig. 11, `ol.data` is comparable to the traditional approaches. These results mean that there is no negative impact caused by the split operation of the *enhanced validation step* in terms of loss and accuracy. Meanwhile, `ol.data` shows higher loss values than Numpy over time (i.e., epoch) in Fig. 10(c). To understand the reason, we observed the processing flow in detail and found that the root cause comes from the behavior of tf.data; it is out of scope of this paper. But, we believe that the gap is negligible because `ol.data` completes the inference model 41.8% faster by using multiple threads.

V. RELATED WORK

There were diverse prior efforts in building elegant and efficient machine learning (ML) services. In this section, we describe two categories of related work: (1) optimizaing resource utilization and (2) offloading techniques for ML models.

Optimizing resource utilization. Recent work focused on the internal behaviors and data flow of inference models that can be optimized to better employ the given hardware resources. tf.data [13] proposes a method to pipeline input data processing in the existing ML frameworks. CheckFreq [14] provides a fine-grained checkpointing scheme that reduces the checkpointing cost in compute-intensive and time-consuming ML models using the pipelined snapshot() and persist() phase. DBS [15] allows to dynamically adjust the batch size and dataset according to the performance of the worker in distributed training of deep neural networks (DNNs). Synergy [16] provides the optimistic profiling and schedules a set of job in multi-tenant clusters; it is a workload and data-aware scheduler. The closest to our work is tf.data, but it only focuses on processing the input data in parallel, unlike `ol.data`. In addition, `ol.data` can bring better resource utilization because it can overlap training steps with validation steps via the *data-level isolation*.

Offloading techniques for ML models. In recent, some researchers have explored optimization opportunities in ML by offloading its techniques to hardware. FlashNeuron [22] offers a peer-to-peer direct path for data transfer between a GPU and NVMe SSDs to reduce the interference of host resources, such as CPU and DRAM. Ren et al. [21] also focus on reducing the amount of data movement between GPU and CPU and they proposed three parallelism levels along with an acceleration system based on Intel high-end FPGAs. FlexLearn [17] proposes a flexible on-chip learning architecture that includes the specialized data paths and dependencies to optimize parallelism for brain simulation.

Finally, we believe even if the previous studies can significantly accelerate the performance of inference models, they are undesirable in diverse environments because of requiring hardware support.

VI. CONCLUSION

Designing a high-performance machine learning (ML) model based on rich hardware resources needs a clear understanding of the internal data flow. In this paper, we have described how the existing inference model works and utilizes CPU and GPU resources. We also proposed `ol.data` for fully utilizing the wasted hardware resources. The key insight in `ol.data` is to (1) reduce the time waits for completing the previous step of the data flow inside the inference model and (2) leverage the strengths of multiple threads in terms of resource utilization. To realize our insight, we enhanced the steps of the data flow with the *data-level isolation* and *enhanced validation step*. By implementing CNN model, we compared the effectiveness of `ol.data` with two traditional approaches, such as Numpy and tf.data. Our experimental results based on three widely used datasets confirmed that the inference model with `ol.data` is faster up to 41.8% on Extended MNIST dataset while guaranteeing high CPU and GPU utilization. Especially, compared to the state-of-the-art approach, `ol.data` can increase CPU and GPU utilization up to 75.7% and 30.2%, respectively. Finally, we hope is that our work will make diverse ML models easily and benefit future developers when designing and implementing the models.

REFERENCES

[1] Q. Zhou, S. Guo, Z. Qu, J. Guo, Z. Xu, J. Zhang, T. Guo, B. Luo, and J. Zhou, "Octo: INT8 Training with Loss-aware Compensation and Backward Quantization for Tiny On-device Learning," in *Proceedings of 2021 USENIX Annual Technical Conference (USENIX ATC 21)*, 2021, pp. 177–191.

[2] M. Abadi, B. Paul, C. Jianmin, C. Zhifeng, D. Andy, D. Jeffrey, D. Matthieu, y. G. Sanja, I. Geoffrey, I. Michael, K. Manjunath, L. Josh, M. Rajat, M. Sherry, G. M. Derek, B. Steiner, T. Paul, V. Vijay, W. Pete, W. Martin, Y. Yuan, and Z. Xiaoqiang, "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[3] J. H. Park, G. Yun, C. M. Yi, N. T. Nguyen, S. Lee, J. Choi, S. H. Noh, and Y. ri Choi, "HetPipe: Enabling Large DNN Training on (Whimpy) Heterogeneous GPU Clusters through Integration of Pipelined Model Parallelism and Data Parallelism," in *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, 2020, pp. 307–321.

[4] "NVIDIA A100 Tensor Core GPU Architecture," https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf, 2020, [Online; accessed 12-August-2021].

[5] "Apple M1," https://www.apple.com/newsroom/2020/11/apple-unleashes-m1/, 2020, [Online; accessed 20-August-2021].

[6] J. et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 2017, p. 1–12.

[7] "Tensorflow," https://www.tensorflow.org/?hl=en, 2020, [Online; accessed 20-August-2021].

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 675—-678.

[10] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015, [Online; accessed 12-August-2021].

[11] U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. S. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, and X. Zhang, "The Architectural Implications of Facebook's DNN-Based Personalized Recommendation," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2020, pp. 488–501.

[12] C. A. Gomez-Uribe and N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," *ACM Transactions on Management Information Systems (TMIS)*, pp. 1–19, 2015.

[13] D. G. Murray, J. Simsa, A. Klimovic, and I. Indyk, "tf.data: A Machine Learning Data Processing Framework," *Very Large Data Bases (VLDB)*, vol. 14, no. 12, p. 2945–2958, 2021.

[14] J. Mohan, A. Phanishayee, and V. Chidambaram, "CheckFreq: Frequent, Fine-Grained DNN Checkpointing," in *Proceedings of 19th USENIX Conference on File and Storage Technologies (FAST 21)*, 2021, pp. 203–216.

[15] Q. Ye, Y. Zhou, M. Shi, and J. Lv, "FLSGD: free local SGD with parallel synchronization," *The Journal of Supercomputing*, pp. 1–24, 2022.

[16] J. Mohan, A. Phanishayee, J. Kulkarni, and V. Chidambaram, "Synergy: Resource Sensitive DNN Scheduling in Multi-Tenant Clusters," *CoRR*, vol. abs/2110.06073, 2021. [Online]. Available: https://arxiv.org/abs/2110.06073

[17] E. Baek, H. Lee, Y. Kim, and J. Kim, "FlexLearn: Fast and Highly Efficient Brain Simulations Using Flexible On-Chip Learning," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 304–318.

[18] H. Jang, J. Kim, J.-E. Jo, J. Lee, and J. Kim, "MnnFast: A Fast and Scalable System Architecture for Memory-Augmented Neural Networks," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, p. 250–263.

[19] J. Liu, H. Zhao, M. A. Ogleari, D. Li, and J. Zhao, "Processing-in-Memory for Energy-Efficient Neural Network Training: A Heterogeneous Approach," in *Proceedings of 2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2018, pp. 655–668.

[20] J. Ren, S. Rajbhandari, R. Y. Aminabadi, O. Ruwase, S. Yang, M. Zhang, D. Li, and Y. He, "ZeRO-Offload: Democratizing Billion-Scale Model Training," in *Proceedings of 2021 USENIX Annual Technical Conference (USENIX ATC 21)*, 2021, pp. 551–564.

[21] D. Kwon, S. Hur, H. Jang, E. Nurvitadhi, and J. Kim, "Scalable Multi-FPGA Acceleration for Large RNNs with Full Parallelism Levels," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[22] J. Bae, J. Lee, Y. Jin, S. Son, S. Kim, H. Jang, T. J. Ham, and J. W. Lee, "FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks," in *Proceedings of 19th USENIX Conference on File and Storage Technologies (FAST 21)*, 2021, pp. 387–401.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "GradientBased Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[25] K. Fukushima and S. Miyake, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition," in *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[26] Y. LeCun, C. Cortes, and C. J. Burges. (2013) The MNIST Database of handwritten digits. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[27] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-MNIST. [Online]. Available: https://github.com/zalandoresearch/fashion-mnist

[28] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. (2017) The EMNIST Dataset. [Online]. Available: https://www.nist.gov/itl/products-and-services/emnist-dataset

[29] S. Godard, *sysstat - System performance tools for the Linux operating system*, Paris, France, 1999. [Online]. Available: http://sebastien.godard.pagesperso-orange.fr/download.html

[30] J. Choi, *gpustat*, Ann Arbor, MI, USA, 2016. [Online]. Available: https://github.com/wookayin/gpustat

JINSEO CHOI received the B.S. degree in the Department of Computer Engineering at Changwon National University in South Korea. He is currently pursuing the M.S. degree in the Department of Computer Engineering with Changwon National University in South Korea. His research interests include operating systems, system software, and system for AI.

DONGHYUN KANG is an assistant professor with the Department of Computer Engineering at Changwon National University in South Korea. Before joining Changwon National University, he was an assistant professor with Dongguk University (2019-2020) and a software engineer at Samsung Electronics in South Korea (2018-2019). He received his Ph.D. degree in College of Information and Communication Engineering from Sungkyunkwan University in 2018. His research interests include file and storage systems, operating systems, system for AI, and emerging storage technologies.