

DACAPO: Accelerating Continuous Learning in Autonomous Systems for Video Analytics

Yoonsung Kim Changhun Oh Jinwoo Hwang Wonung Kim Seongryong Oh
 Yubin Lee Hardik Sharma^{†*} Amir Yazdanbakhsh[‡] Jongse Park

KAIST [†]Meta [‡]Google DeepMind

{yskim, choh, jwhwang, wukim, sroh, yblee}@casys.kaist.ac.kr

hardiksharma@meta.com ayazdan@google.com jspark@casys.kaist.ac.kr

Abstract—Deep neural network (DNN) video analytics is crucial for autonomous systems such as self-driving vehicles, unmanned aerial vehicles (UAVs), and security robots. However, real-world deployment faces challenges due to their limited computational resources and battery power. To tackle these challenges, continuous learning exploits a lightweight “student” model at deployment (*inference*), leverages a larger “teacher” model for labeling sampled data (*labeling*), and continuously retrains the student model to adapt to changing scenarios (*retraining*). This paper highlights the limitations in state-of-the-art continuous learning systems: (1) they focus on computations for retraining, while overlooking the compute needs for inference and labeling, (2) they rely on power-hungry GPUs, unsuitable for battery-operated autonomous systems, and (3) they are located on a remote centralized server, intended for multi-tenant scenarios, again unsuitable for autonomous systems due to privacy, network availability, and latency concerns. We propose a hardware-algorithm co-designed solution for continuous learning, DACAPO, that enables autonomous systems to perform concurrent executions of inference, labeling, and retraining in a performant and energy-efficient manner. DACAPO comprises (1) a spatially-partitionable and precision-flexible accelerator enabling parallel execution of kernels on sub-accelerators at their respective precisions, and (2) a spatiotemporal resource allocation algorithm that strategically navigates the resource-accuracy tradeoff space, facilitating optimal decisions for resource allocation to achieve maximal accuracy. Our evaluation shows that DACAPO achieves 6.5% and 5.5% higher accuracy than a state-of-the-art GPU-based continuous learning systems, Ekya and EOMU, respectively, while consuming 254× less power.

I. INTRODUCTION

Live video analytics is key to enable seamless functionality of interactive autonomous systems, such as autonomous driving vehicles [10], [25], [35], [38], [42], [87], [93], unmanned aerial vehicles (UAVs) [55], [56], [59], [60], [101], and security surveillance robots [7]–[9], [24], [71], [79], [92]. Real-time analysis of live videos is indispensable for comprehending rapidly changing environmental conditions, a pivotal aspect that ensures the successful execution of the autonomous systems’ missions.

Modern video analytics heavily relies on the capabilities of deep neural networks (DNNs). However, the computational demand to execute DNN inference on individual video frames

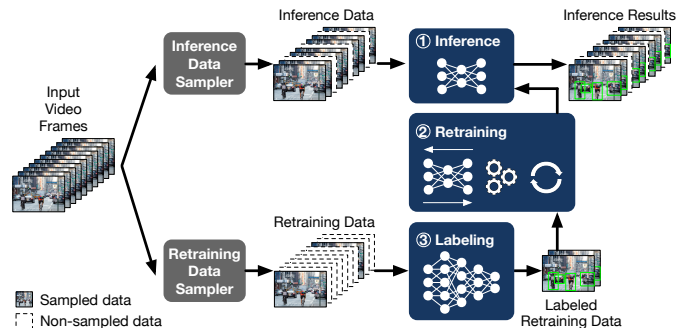


Fig. 1: Overview of continuously learning video analytics in autonomous systems. To address privacy, networking cost, and latency concerns, autonomous systems exclusively use constrained computing resources to concurrently execute the three continuous learning kernels – (1) inference, (2) retraining, and (3) labeling – which presents a performance challenge.

is far beyond the capabilities of real-world systems. Although sampling can decrease computational demands, it often compromises accuracy, especially in videos captured by moving cameras, a frequent condition in autonomous systems.

To address this issue, *continuous learning* (CL), an emerging algorithmic methodology, has gained significant attention in recent years [5], [27], [34], [46]. This innovative approach enables autonomous systems to perform inference using lightweight “student” DNN models, which results in reduced resource consumption. However, the limited generality of the lightweight student models gives rise to the *data drift* problem, where live video data deviates from training data. Continuous learning incorporates constant *on-the-spot* retrains, which introduce a new computational challenge, as the systems must now handle not only *inference*, but also *labeling* and *retraining*, as illustrated in Figure 1. We denote these three tasks as *continuous learning kernels* in this paper.

Two recent pioneering works, Ekya [5] and RECL [46], propose initial system designs for continuously learning video analytics. Although these works have paved the way for a new research direction, their direct deployment to autonomous systems is infeasible due to the following reasons: (1) these systems overlook the necessity of computation for label-

*This work was done at Google.

ing and inference, a factor that could be substantial for battery-operated, resource-constrained autonomous systems; (2) they assume excessively rich computation resources for autonomous systems; and (3) they are designed to be deployed on a remote centralized server, intended for multi-tenant scenarios. For instance, Ekya [5] runs on a multi-GPU edge server and RECL [46] offloads the retraining computation to the scalable cloud, both largely focusing on the retraining computations of multiple models.

In this paper, we propose a hardware-software co-designed acceleration solution, DACAPO, that overcomes the limitations of existing systems and enables effective deployment of continuously learning video analytics on autonomous systems without the need for remote access to abundant resources. In designing DACAPO, we identify the following challenges:

- **Challenge 1:** DACAPO must offer significantly higher performance and energy efficiency compared to the existing high-performance computing platforms such as GPUs, to effectively implement continuous learning for video analytics on battery-operated autonomous systems.
- **Challenge 2:** DACAPO requires a flexible and dynamically reconfigurable accelerator architecture to rapidly adapt to the changing resource demands of the three continuous learning kernels when data drifts occur.
- **Challenge 3:** DACAPO must be able to capture the resource-accuracy tradeoff of the continuous learning kernels and efficiently utilize the computing resources to maximize the resulting accuracy.

To address the aforementioned challenges, we make the following contributions.

- **Spatially-partitionable and precision-flexible accelerator architecture.** We architect the DACAPO accelerator, employing a spatially-partitionable architecture, which can be dynamically composed into sub-accelerators (SAs). The design principle aims to attain flexibility for better resource utilization without imposing unnecessarily large overhead to enable this adaptability. To this end, the DACAPO accelerator is organized as vertically-stacked rows of processing elements in which the rows can be partitioned and grouped into two sub-accelerators. A processing element in DACAPO performs a vector dot product rather than a multiplication, which we call Dot-Product Engine (DPE), leveraging recently proposed MX-based precision-flexible arithmetic to unleash higher performance and energy-efficiency from the accelerator. This spatial sharing and precision configurability are the keys that enable DACAPO to obtain orders of magnitude higher efficiency and allow the three kernels to fully utilize the given resources for maximal accuracy.
- **Spatiotemporal resource allocation algorithm for maximal accuracy.** Resource requirements for the three kernels vary depending on the runtime scenarios that constantly change. A key insight driving our resource allocation algorithm is that *labeling* plays a pivotal role in the occurrence of data drifts for promptly collecting retraining data with new data patterns, requiring relatively

more computing resources. To maximize accuracy using the limited resources, we devise a spatiotemporal resource allocation algorithm that makes static (spatial) and dynamic (temporal) decisions for resource allocations. First, at offline, the algorithm statically partitions the accelerator into two sub-accelerators, T-SA and B-SA, based on the resource requirement of inference to keep up with the input frame rate. Then, during runtime, the algorithm dynamically detects data drifts through constant accuracy validations and when it is detected, the algorithm allocates more time slots for labeling until the dataset is sufficiently updated (temporal decision).

To evaluate DACAPO’s effectiveness, we use three pairs of (student, teacher) object recognition models (see Table III) and the BDD100K driving video dataset. We model three types of data drift: (1) task change by addition or removal of object labels, (2) transitions between daytime and nighttime, and (3) location change between city and highway. These data drifts are mixed to synthesize size different workload scenarios representing diverse real-world situations that autonomous systems would face. We develop DACAPO using Verilog RTL and synthesize it using Synopsys Design Compiler with TSMC 28nm technology. To cross-validate the cycle-accurate behaviors of DACAPO, we also develop an in-house simulator built upon SCALE-Sim [77].

Our evaluation shows that DACAPO achieves 6.5% and 5.5% higher accuracy than the two state-of-the-art GPU-based continuous learning systems, Ekya and EOMU, respectively. Furthermore, we observe that DACAPO achieves the accuracy improvement, consuming $254\times$ less energy compared to the GPU baseline. These advantages highlight DACAPO’s role in connecting the algorithmic advancements in continuous learning to their practical deployment on autonomous systems. Our tools, including system and hardware simulators and a continuous learning benchmark suite with datasets, are available at <https://github.com/casys-kaist/DaCapo>.

II. BACKGROUND

This paper aims to develop an acceleration solution for continuously learning video analytics on autonomous systems. In this section, we briefly outline the target problem and introduce existing solutions.

A. Video Analytics in Autonomous Systems

Modern autonomous systems leverage compute-heavy deep neural network (DNN) models for complex video analytics. Constrained by limited resources, these systems struggle to meet latency requirements, often requiring compromises that can entail frame drops and a potential reduction in overall analytical quality. Recently, using less capable yet lightweight DNN models has been introduced as a viable solution [4], [36], [37], [48], allowing the systems to match the inference rate with the frame rate. However, the customized models present a new challenge known as *data drift*, where live video data diverges from the training data, consequently reducing the accuracy of the lightweight models.

B. Continuously Learning for Video Analytics at Edge

To address the data drift problem, a large body of literature [5], [46], [82], [85] has explored *continuous learning*, which aims to continuously adapt models to newly appearing data. This approach allows the models to stay up-to-date with the changing data distribution. While continuous learning is a promising approach to reduce compute demand for inference, it introduces new compute load for constant retraining over newly collected data as well as labeling the retraining data by running inference for large, state-of-the-art models, often referred to as *teacher* models [66]. The challenge arises from the necessity of executing the three distinct yet compute-intensive kernels, each affecting the final accuracy differently, creating a complicated tradeoff space.

C. Continuous Learning Systems for Video Analytics

Inspiring prior works, Ekya [5], and RECL [46], are pioneering efforts that address the challenge, approaching it from a system design perspective. Ekya and RECL offer resource allocation and scheduling solutions tailored to GPU-accelerated live video analytics platforms. In particular, they address the “at-scale” problem, where the systems are required to manage and process multiple DNN models specific to various video streams concurrently. Optimizing for multi-tenant workload scenarios, the existing systems are centered on facilitating resource sharing across different models, requiring the use of mid-size to cloud-scale servers equipped with multiple compute-powerful computing platforms such as datacenter-level GPUs. A recent work, EOMU [49], takes a similar yet different approach by proposing a log-based adaptive resource allocation technique for edge-cloud continuous learning system. In contrast, this work aims to focus on enabling single-task video analytics on resource-constrained autonomous systems running the three continuous learning kernels, a distinctive and novel challenge.

III. CHALLENGES AND OPPORTUNITIES

In this section, we examine the limitations of current solutions, identify key challenges, and explore potential opportunities for improvement.

A. Unveiling the Dilemma in Continuous Learning

To understand the performance implications, we perform a preliminary experimental study, using two pairs of Student and Teacher object recognition models: (S:ResNet18, T:WideResNet50) and (S:ResNet34, T:WideResNet101). For the experiments, we use an NVIDIA 3090 and a Jetson Orin’s GPU, representing datacenter-scale and autonomous system-level GPUs respectively. As a baseline for continuous learning systems, we built Ekya under an idealized assumption that the system can optimally utilize the provided GPU resources to obtain the maximum achievable accuracy.

Figure 2 reports the accuracy results of student/teacher models compared to the Ekya baseline, measured by calculating the ratio of correct predictions to the total number of classifications. Dropped frames produced by insufficient computing

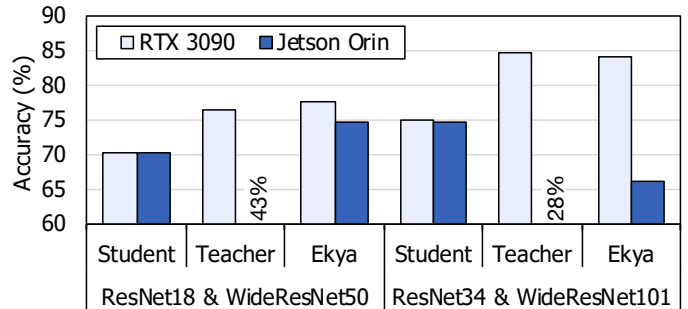


Fig. 2: Accuracy comparisons of Ekya versus student and teacher models. Student and teacher models are *non-continuous learning* cases. The experiments are conducted on RTX 3090 and Jetson Orin. The accuracy gap between the two GPUs is attributed to inevitable frame drops due to a lack of computing resources.

resources are considered as incorrect results. The RTX 3090 has sufficient compute resources, resulting in no frame drops. Therefore, the results are solely evaluated by their algorithmic performance. The teacher models always substantially outperform student models, exhibiting their algorithmic superiority. While Ekya employs the student model for inference, its accuracy often approaches or even exceeds that of the teacher model since Ekya can adeptly customize the student model to suit the specific system environment. However, as we change the compute platform from RTX 3090 to Orin, for both teacher and Ekya baselines we observe significant accuracy drops. These results demonstrate that continuous learning effectively mitigates computational bottlenecks, but still notable accuracy loss persists when insufficient resources are provided, the key challenge addressed in this work.

Existing continuous learning systems function effectively only with substantial computational resources, a condition incompatible with the limitations of autonomous systems. This insight underscores the necessity for specialized hardware innovation, crucial for obtaining levels of performance and efficiency otherwise unachievable, thereby unleashing the full potential of continuous learning in autonomous systems.

B. Workload Characterization of the Three Kernels

A straightforward solution for the aforementioned challenge is to develop a high-performance accelerator with ample compute resources and exploit time-multiplexing to flexibly schedule the three kernels interchangeably on the chip. However, such a time-sharing technique would limit resource utilization since the three kernels offer different levels of parallelism (e.g., inference offers much lower data parallelism than retraining). Alternatively, the three kernels can be spatially parallelized by designing an accelerator with three separate sub-modules, with each module dedicated to a specific kernel. We found that this approach is also suboptimal since the computational demands of the kernels exhibit temporal variations at runtime.

To understand the workload of continuous learning, we explore the computational demands of running a continuous

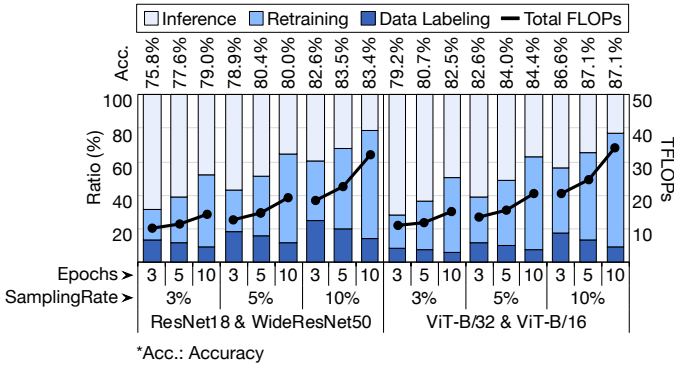


Fig. 3: MAC operation breakdown of the three kernels and total FLOPs for the entire experiment runs.

learning system for 120 seconds, measuring these in FLOPs, and examine how each of the three kernels contributes to the overall FLOPs count. Figure 3 presents the results, with the stacked bars illustrating the proportional distribution of FLOPs for inference, retraining, and labeling from top to bottom, while the line graph delineates how the total FLOPs scale over time. We observe a notable shift in proportions: the proportion allocated to retraining surges from 26.0% to 82.3% as data sampling rates for labeling and number of epochs for retraining increase, while inference and labeling decrease from 57.8% and 27.1% to 9.1% and 7.0% respectively. The results also show that increasing the sampling rate and number of epochs improves accuracy, albeit at the expense of increased FLOPs.

The overall accuracy of a continuous learning system is jointly influenced by the interplay of the three kernels, whose contributions shift based on hyperparameters, leading to varying computational demands. Inspired by this insight, we design the DACAPO accelerator to enable dynamic and flexible resource allocation among the kernels. This approach aims to optimize resource efficiency, consequently improving the final accuracy.

C. Opportunities from Low-Precision Arithmetics

As discussed earlier, effective continuous learning in an autonomous system demands not just exceptional performance but also significant energy efficiency, which motivates the exploration of strategies that can achieve enhanced performance and efficiency, without sacrificing accuracy. Low-precision arithmetic through quantization has been widely adopted to significantly reduce the computational resource demands for training and inference [19], [23], [50], [73]. Among various quantization formats, block floating point (BFP) has recently gained prominence owing to its hardware-friendly characteristics and ability to support a wide range of real values [19], [23], [73]. BFP groups a set of floating point values, forces them to have a shared exponent by shifting the mantissa accordingly, and stores the group of truncated mantissa bits along with the shared exponent. This way, most computations happen in the integer domain, which is significantly cheaper than single-precision floating-point (FP32) arithmetics, while still offering FP32-like algorithmic behaviors.

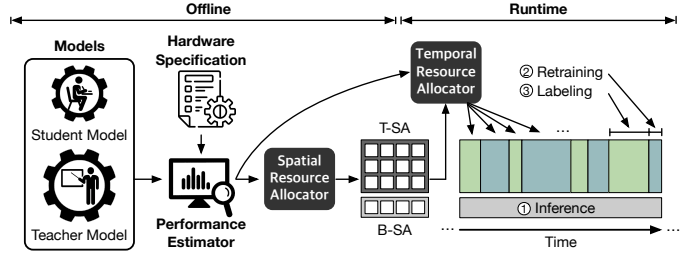


Fig. 4: Workflow of a DACAPO-based continuously learning video analytics system.

Particularly, a recent work from Microsoft [19] proposes a customized block floating point format, called MX (MicroXponent), and demonstrates the effectiveness on both training and inference. The prior work also suggests that in general, training requires higher precision (MX9) than inference (MX6 or MX4), and the best-fit MX precision varies depending on the particular models. However, the MX accelerator is not precision-flexible, lacking support for multiple precisions on a single hardware, which is not well-suited for continuous learning scenarios where an accelerator should simultaneously perform both training and inference for different models. Therefore, to effectively employ MX and fully realize the potential for continuous learning, there is a need for customized architectural techniques, which is one of the goals of this work.

While MX offers promising opportunities, the accelerator needs to support various MX precisions, switchable at runtime. This flexibility is essential to fully capitalize on the advantages of mixed computations in both inference and retraining executions across the three kernels.

IV. OVERVIEW OF DACAPO'S SYSTEM WORKFLOW

Figure 4 illustrates the overall workflow of DACAPO-accelerated continuous learning in an autonomous system.

1 System initialization. The system needs to be initialized with the student and teacher DNN models for inference and labeling. These models are pre-trained over the general dataset without having any specific context that the system is actually used for. We assume that the hyperparameters and optimization techniques for on-the-fly retraining are pre-determined and not adjusted at runtime. Additionally, we also assume that the frame rate of input video and its resolution are a priori known knowledge and never change at runtime.

2 Performance estimation. The subsequent stage involves our simulation-based performance estimator, which utilizes the architecture of the specified student/teacher models and the hardware specifications of the DACAPO accelerator to approximate the throughput of the three kernels. In this phase, the estimator examines every supported MX precision, ranging from MX4 through MX6 to MX9, to assess its impact on accuracy. Our observations show that the models evaluated achieve stable outcomes when using MX9 for retraining and MX6 for inference/labeling, whereas utilizing MX4 leads to considerable accuracy degradation in both scenarios. Note that this observation aligns with the experimental findings reported

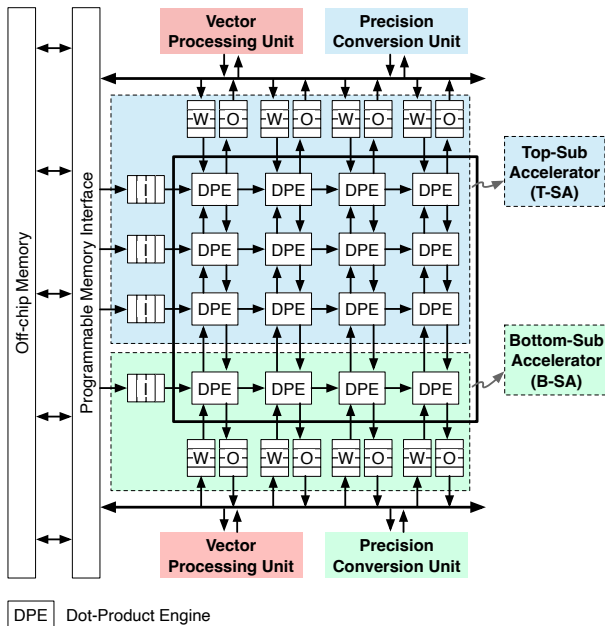


Fig. 5: Overall architecture of DACAPO.

in the original MX paper [19]. This crucial information is then relayed to the resource allocators in subsequent stages, aiding them in their decision-making processes.

3 Offline spatial resource allocation. Utilizing the performance estimation, our spatial resource allocator calculates the minimum number of rows needed for B-SA to process inferences for input frames, ensuring it matches the input frame rate. Although the MX precisions remain constant during runtime, precision flexibility is still necessary because a row may be allocated to either T-SA or B-SA based on the decisions made in spatial resource allocation.

4 Online temporal resource allocation. While B-SA consistently handles inference computations, T-SA alternates between retraining and labeling tasks through time-sharing. In formulating the temporal resource allocation algorithm, we exploit an empirical insight that when data drifts occur, rapidly labeling data with a new distribution is more crucial than conducting additional retraining iterations. For DACAPO to respond effectively to data drifts, it must first detect them. To achieve this, DACAPO tracks accuracy trends in real-time, employing a small portion of the labeled training dataset as a validation set for identifying any data drifts. Upon their detection, DACAPO adjusts its time allocation, favoring labeling over retraining. This shift in temporal resources aids the system in quickly recovering from accuracy losses caused by data drifts. The empirical evidence supporting this approach will be detailed in Section VII.

V. DACAPO ACCELERATOR ARCHITECTURE

Figure 5 presents the overall architecture of the DACAPO accelerator. The design is built upon a standard systolic array architecture, which is composed of two-dimensional processing engines. There are two primary differences compared to conventional TPU-like systolic array architectures:

(1) the accelerator is *spatially* partitionable into Top Sub-Accelerator (T-SA) and Bottom Sub-Accelerator (B-SA); and (2) the processing engines are not simple multiplication-and-accumulation (MAC) units, but dot-product units that can perform vector dot products of 2-bit, 4-bit, and 8-bit operands, depending on the MX precision mode. This section will describe the two architectural techniques in more detail.

A. Spatially-Partitionable Architecture

Why revisit spatially partitionable architectures? Spatial partitioning techniques for neural network accelerators have been explored in previous works, such as Planaria [26] and Dataflow mirroring [52]. Planaria utilizes architectural fission to dynamically compose processing elements into an arbitrary number of sub-accelerators, enabling efficient multi-DNN inference and adherence to service-level objectives (SLOs). However, Planaria’s on-chip buffer partitioning and inter-buffer networks are quite complex. Dataflow mirroring proposes a simpler four-way partitioning scheme for systolic arrays, prioritizing efficiency over flexibility. While it provides simpler on-chip memory and networking architecture, it targets multi-tenant inference serving, which demands excessive adaptability than continuous learning.

To address this, we introduce a two-way spatial partitioning approach that effectively balances architectural simplicity with the necessary flexibility, efficiently facilitating the parallel execution of the three kernels.

Row-granular processing element organization. For simplicity, we propose to group the two-dimensional processing elements of systolic arrays into one-dimensional *rows*, which can be composed into two groups: T-SA and B-SA. In typical systolic arrays, the input, weight, and output tensors flow either horizontally or vertically, depending on which data is stationary. In designing the dataflow for DACAPO, leveraging the property that both weights and outputs flow vertically, we propose to enable vertical communication channels in “both” directions, placing weight and output buffers at both the top and bottom of the systolic array. This way, regardless of how the rows are assigned to T-SA and B-SA, weights and outputs can flow vertically, so that the two stacked SAs can simultaneously run independent matrix-matrix multiplications without interfering with the other section. While we can support both weight and output stationary designs, we employ output stationary design for our RTL prototype.

Programmable memory interface. The proposed spatial partitioning approach requires T-SA and B-SA to place tensors at the input (I) and weight (W) buffers differently, as their dataflow differs. We do not perform this tensor transformation in software, but a programmable memory interface orchestrates the memory layout. Once our resource allocation algorithm determines the row assignments for T-SA and B-SA, it also reprograms the memory interface to feed the operands into the on-chip buffers in a way that the sub-SAs can directly read. The memory controller is designed to support MX and tensor manipulation operations using MX metadata

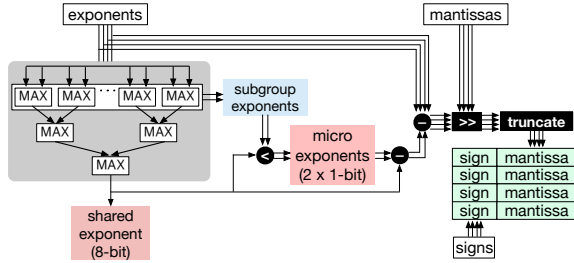


Fig. 6: Conversion of four FP32 values into MX format.

decoder and bitwise concatenator to produce DPE operands. Note that this programmable memory interface is largely based on the memory interface used by several prior work, including MX [19], FAST [73], and BitFusion [80].

B. Precision-Flexible Dot-Product Engine

Another dimension of reconfiguration in the DACAPO accelerator is MX precision [19]. While it is statically determined for each kernel, every processing element must support multiple MX precisions, since they can be assigned to different SAs. We describe the microarchitectural design of these processing elements, which we call Dot-Product Engine (DPE).

MicroExponent (MX) format. Figure 6 illustrates how MX compresses a block of values differently from the conventional block floating point format. As in regular block floating point, MX compresses 16 address-adjacent values in memory as a block. While the choice of block size is orthogonal to our DPE design, we use 16 as it is commonly used in most prior works [19], [67], [73]. However, for clearer illustration purposes, Figure 6 employs a block size of 4. MX first identifies the largest exponent within a block using a max tree and chooses it as the shared exponent. Each block consists of sub-blocks where its values share a 1-bit microexponent (MX). The sub-block size can vary, but we use 2, the same default value used in the original paper. MX compares the maximum subgroup exponents obtained from the leaf stage of max tree with the shared exponent and sets the MX bit to 1 if every subgroup exponent is smaller than the shared exponent, effectively enlarging the supported dynamic range and thus enhancing accuracy resilience in lossy compression. After determining the shared exponent and microexponents, mantissa values are shifted accordingly and at last, truncated from 23-bit of single-precision floating point to 2 (MX4), 4 (MX6), or 7 bits (MX9).

Microarchitectural opportunities from prior works. As the name implies, DPEs perform vector dot products between two MX-compressed vectors. The DPE microarchitecture is built upon techniques proposed in two prior works, Bit Fusion [80] and SIGMA [74]. Bit Fusion [80] proposes bit-parallel dynamic fusion of small-bit width multiplications for quantized neural network inference. SIGMA [74], on the other hand, proposes a forwarding adder network (FAN) that effectively reduces the results of multiple dot products using an adder tree structure equipped with a forwarding datapath. Inspired

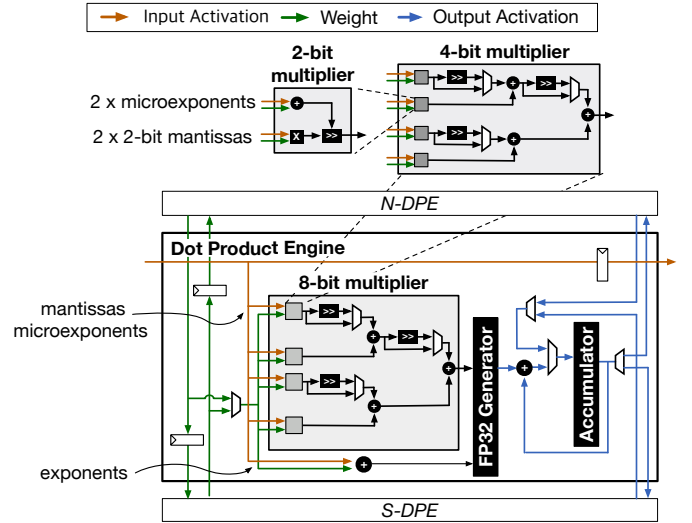


Fig. 7: Microarchitecture of Dot-Product Engine (DPE).

by these works, we leverage the proposed techniques in prior works for our DPE design, customizing them for performing dot products between MX-compressed vectors.

Dot-product engine: a hierarchical architecture for precision-flexible dot-product. Figure 7 illustrates the microarchitectural details of a DPE. To support various MX formats, we employ a hierarchical multiplication-and-accumulation (MAC) tree that can support 16×2 -bit (MX4), 4×4 -bit (MX6), or 1×8 -bit (MX9) in parallel. While the different MX modes require different numbers of mantissa values, the memory interface partitions the mantissa values in the granularity of 2-bits and supplies an array of 2-bits to the corresponding 2-bit multipliers. When it runs in the MX4 mode, every 2-bit multiplier performs independent multiplications, while for MX6 and MX9 mode, the results of 2-bit multipliers are accumulated by the hierarchical adder trees, producing 4-bit and 8-bit multiplication results, respectively.

Intra-DPE forwarding datapath and FP32 generator. Note that the multiplication results must be grouped in a specific size (e.g., 16) and then compressed into the MX format. Therefore, the accumulated results from the hierarchical MAC tree need to be collected at the FP32 generator, which requires direct result bypassing. To facilitate this, we devise a hierarchical result forwarding datapath, which can send multiplication outcomes through the MAC tree to the FP32 generator from any multiplier level. The rate of FP32 result generation depends on the MX format. For instance, MX4 completes all multiplications and accumulations for a dot product in just one cycle, whereas MX6 and MX9 require 4 and 16 cycles, respectively, due to the serial processing of subsets of the 16 multiplications (i.e., MX9: 16 cycles \times 1 multiplication and MX6: 4 cycles \times 4 multiplications). After the FP32 generator finishes the accumulations, it converts the final result into a floating-point value.

Inter-DPE communication. Each DPE has bidirectional inter-

DPE communication channels to its north and south. It also has a horizontal channel, which pipelines the input activations from west to east, as in existing systolic array designs. The two vertical channels are activated simultaneously. One streams weight parameters to DPEs, while the other drains accumulated results to output buffers.

C. Precision-Conversion Unit

Note that when a set of results are produced by a SA and received by the output, they are floating-point values, not compressed in MX format. Therefore, the DACAPO accelerator comes with two precision-conversion units for the two SAs, which partition the floating-point output values into MX blocks of size 16. For inference and labeling, only row-major matrices are produced by the precision-conversion unit. However, in the case of retraining, the precision-conversion unit carries out column-major MX conversion, in addition to the default row-major conversion, because transposed matrices are needed for gradient computation and weight updates.

VI. SPATIOTEMPORAL RESOURCE ALLOCATION ALGORITHM

This section describes DACAPO’s resource allocation algorithm, which is pivotal for optimizing continuous learning.

A. Algorithm Overview

Our resource allocation algorithm operates on both spatial and temporal dimensions, determining spatial allocations statically during the offline phase and continuously adjusting temporal allocations dynamically in the online phase. Unlike conventional fixed-window periodic retraining methods, DACAPO’s temporal resource allocation mechanism promptly alternates between the retraining and labeling phases, with their durations governed by the algorithm-determined number of samples for each cycle. At the end of each cycle, DACAPO assesses the impact of these phases on accuracy and adjusts temporal resource allocation dynamically.

This algorithm requires maintaining a fixed-capacity sample buffer to store labeled samples. For retraining, a subset is drawn from this buffer, which is subsequently updated with newly labeled data in the following labeling phase. At the end of retraining, DACAPO compares the validation accuracy of existing data with that of newly labeled data to detect data drift. This facilitates DACAPO to make informed decisions regarding resource allocation for labeling in upcoming phases. In the following section, we provide a formal description of the algorithm’s operation.

B. Hyperparameters for Resource Allocation

Table I shows the hyperparameters relevant to spatiotemporal resource allocation. They are primarily concerned with the number of samples and influence the duration of retraining and labeling phases. As DACAPO operates without a fixed window time, the values assigned to these hyperparameters dictate the scheduling of the retraining and labeling phases.

Hyperparameters for retraining. For retraining, DACAPO loads a subset of samples from the sample buffer with buffer

TABLE I: Notations of hyperparameters used in DACAPO’s spatiotemporal resource allocation algorithm.

Retraining Hyperparameters	
N_t	Number of samples for retraining
N_v	Number of samples for validation
Labeling Hyperparameters	
N_l	Number of samples to label at usual
$N_{l_{dd}}$	Number of samples to label at data drift
Spatial Allocation Hyperparameters	
R_{tsa}	Number of DPE rows at T-SA
R_{bsa}	Number of DPE rows at B-SA

capacity C_b . The number of samples for retraining (N_t) and validation (N_v) are predefined, while N_v is set to half of N_t . These hyperparameters are decided according to the model size, as it has a direct impact on the computational cost required for retraining.

Hyperparameters for labeling. The quantity of samples for labeling is specified by two distinct hyperparameters: N_l and $N_{l_{dd}}$. N_l specifies the standard number of samples to be labeled under normal conditions, while $N_{l_{dd}}$ is an increased number of samples to be selected for labeling when data drift is detected. In DACAPO, $N_{l_{dd}}$ is established as a multiple of N_l . We empirically select this multiple to be four.

Hyperparameters for spatial allocation. The number of DPE rows at T-SA and B-SA are defined as R_{tsa} and R_{bsa} . To optimize them, we prioritize the R_{tsa} where the retraining and labeling phases run on, while ensuring that the R_{bsa} , the inference resources, is sufficient to meet the latency requirements of streaming input frames.

C. Spatiotemporal Resource Allocation Algorithm

Algorithm 1 describes the temporal resource allocation algorithm in detail. *Line 1-2*: DACAPO allocates resources for T-SA and B-SA during the offline phase. *Line 4-7*: In the retraining phase, DACAPO evaluates the updated model’s weights (W_{next}) using a validation dataset (D_v) sampled from the current sample buffer (B_{cur}). Note that as validation only involves forward-pass processing on T-SA, its computational cost is significantly lower than that of retraining. The validation accuracy assesses the new model’s adaptability to the accumulated dataset. *Line 8-10*: During the labeling phase, DACAPO matches inference outputs with corresponding samples to evaluate the current model’s inference accuracy on the newly sampled dataset. *Line 11*: If the validation accuracy subtracted from the labeling accuracy is lower than the threshold (V_{thr}), it indicates data drift. *Line 12-13*: Upon detecting data drift, DACAPO clears the current sample buffer to remove outdated samples, and labeling time is extended to ensure that these better-suited samples are included in the dataset for subsequent phases. *Line 14*: Once all new samples are labeled, DACAPO updates its sample buffer. With this algorithm, DACAPO adeptly detects and addresses data drift by dynamically allocating labeling time based on the validation accuracy from a segment of the sample buffer.

Algorithm 1: Spatiotemporal Resource Allocation

Input: B_{cur} : Current sample buffer
 W_{cur} : Current weight of student
 C_b : Capacity of sample buffer
 V_{thr} : Threshold value to detect data drift
 $N_{t,v}$: Retraining hyperparameters
 $N_{l,ldd}$: Labeling hyperparameters

// Spatial resource allocation

- 1 $R_{tsa}, R_{bsa} \leftarrow \text{GetSpatialAllocation}()$;
- 2 $\text{SetAccelerator}(R_{tsa}, R_{bsa})$;

// Temporal resource allocation

- 3 **while** *true* **do**
 - // Retraining
 - 4 $D_t, D_v \leftarrow \text{GetData}(B_{cur}, N_t, N_v)$;
 - 5 $W_{next} \leftarrow \text{Retrain}(W_{cur}, D_t)$;
 - 6 $W_{cur} \leftarrow \text{UpdateWeight}(W_{next})$;
 - 7 $acc_{val} \leftarrow \text{Valid}(W_{cur}, D_v)$;
 - // Labeling
 - 8 $D_l \leftarrow \text{Label}(N_l)$;
 - 9 $O_{prd} \leftarrow \text{GetInferOutputs}(D_l)$;
 - 10 $acc_{lab} \leftarrow \text{Evaluate}(O_{prd}, D_l)$;
 - // Check data drift & update Buffer
 - 11 **if** $acc_{lab} - acc_{val} < V_{thr}$ **then**
 - 12 $\text{ResetBuffer}(B_{cur})$;
 - 13 $D_l \leftarrow \text{Label}(N_{ldd} - N_l) + D_l$;
 - 14 $B_{cur} \leftarrow \text{UpdateBuffer}(B_{cur}, D_l, C_b)$;

TABLE II: Descriptions of workload scenarios for continuously learning video analytics on autonomous systems.

Name	Weather	Data Drift Types (Only 1)
S1	Clear	Label Distribution
S2	Overcast	
S3	Clear	Label Distribution, Time of Day
S4	Snowy	
S5	Clear	Label Distribution, Time of Day, Location
S6	Rainy	
Name	Composition of Data Drift Types (All 4)	
ES1	Label Distribution, Time of Day, Location, Weather	
ES2		

D. Hyperparameter Tuning for Different Environments

As noted in Section VI-B, we use a fixed set of empirically determined hyperparameters for the resource allocation algorithm. To determine the hyperparameters, we use the BDD100K dataset (see Section VII-A) and exhaustively explore the hyperparameter search space to establish their values, consistently finding that the chosen settings outperform alternatives across various environmental scenarios, suggesting a robust insensitivity to scenario changes. Given this empirical insight, we choose to use an offline hyperparameter tuning method, leaving the online hyperparameter auto-tuning to future work. Before deploying autonomous systems in real-

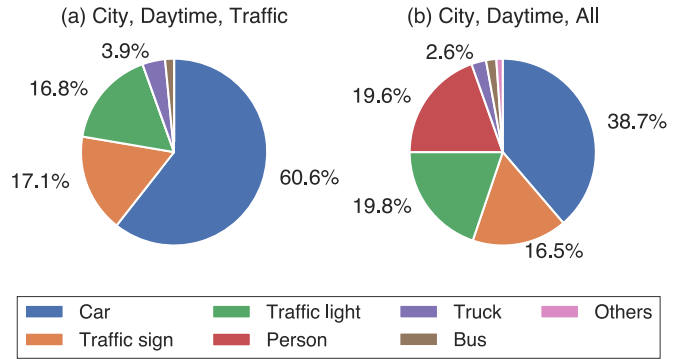


Fig. 8: Different label distributions within distinct 60-second segments of an example scenario.

world environments, DACAPO requires offline hyperparameter tuning to identify the optimal settings. The tuning is required only once for each autonomous system, representing a cost that is not only inexpensive but also amortizable over time.

VII. EVALUATION**A. Methodology**

Dataset and data drifts. We use BDD100K [97], a comprehensive live video driving dataset with an abundance of varied situations. To adapt it for classification, we crop individual objects within the video frames and arrange them in chronological order. To create data drifts, we characterize videos based on three key attributes: Label Distribution, Time of Day, and Location. There are two types of “Label Distribution”: (1) *Traffic Only* which focuses on traffic-related labels, and (2) *All* which includes a broader range of labels, such as pedestrian, bicycle, and motorcycle. “Time of Day” is classified into *Daytime* and *Night* to account for varying lighting conditions. “Location” is divided into *City* and *Highway*, each presenting unique driving environments. This setup facilitates the evaluation of DACAPO’s performance in various data drift situations, providing insights into its flexibility and adaptability.

Scenarios. Every scenario constructed for evaluation is created by concatenating a series of video clips. Each video is treated as a collection of frames within the scenario, unfolding over a duration of 20 minutes, at a frame rate of 30 FPS. We develop three distinct types to capture diverse data drift scenarios including (1) 6 regular scenarios with single data drift at a time (S1–S6), and (2) 2 extreme scenarios with 4 data drifts occurring at the same time (ES1 and ES2). Table II provides the details of 8 scenarios and Figure 8 shows the label distributions present across different segments of scenarios.

DNN models. We evaluate DACAPO’s performance on object classification tasks using six DNN models listed in Table III. These models are paired as student and teacher models: (1) ResNet18 and WideResNet50, (2) ViT-B/32 and ViT-B/16, (3) ResNet34 and WideResNet101.

Accuracy metric. DACAPO differs from other baselines as it does not utilize a window time. For a fair comparison, we

TABLE III: Specifications of the evaluated DNN models.

Type	Name	Parameters	GFLOPs
Student	ResNet18	11.7M	1.82
	ResNet34	21.8M	3.67
	ViT-B/32	88.2M	4.37
Teacher	WideResNet50	68.9M	11.43
	ViT-B/16	86.6M	16.87
	WideResNet101	126.9M	22.80

TABLE IV: Evaluated GPU and DCAPO platforms.

Device	Jetson Orin	DCAPO
Technology	8 nm	28 nm
Area	N/A	2.501 mm ²
Frequency	1.3 GHz	500 MHz
Power	15 - 60 W	0.236 W
DRAM	LPDDR5	LPDDR5
Bandwidth	204.8 GB/s	204.8 GB/s

assess its averaged accuracy in the scenarios over time slices corresponding to the window period.

Retraining and inference hyperparameters. The learning rate of retraining is set to 10^{-3} , and we use SGD optimization with batch size of 16. For inference, we use batch size 1. While the DCAPO hardware supports MX4, MX6, and MX9 precision modes, we use MX9 (7-bit mantissas) for retraining and MX6 (4-bit mantissas) for inference.

Hardware development and synthesis. Our DCAPO prototype constitutes of a systolic array with 16×16 DPEs and a 96KB on-chip SRAM. Note that while our prototype is designed as a tiny chip, DCAPO could scale the number of DPEs to larger configurations (e.g., 32×32) or multiple DCAPO chiplets could be packaged together if there is a need. For evaluation, we implement Verilog RTL design, and verify the results through an RTL simulator. We synthesize in 28nm CMOS technology using Synopsys Design Compiler and CACTI, which provide the dynamic/static power and area of each component of DCAPO, as reported in Table IV. We also develop an in-house cycle-accurate software simulator to measure the execution cycles of DPE cores.

GPU. We evaluate DCAPO against NVIDIA Jetson Orin, a GPU commonly used in autonomous system computing environments, specifications of which are provided in Table IV. Out of the multiple power options provided, we use two options: high power (default power settings) and low power (30W power constraints) in our evaluations, to understand the behavior of our workload on low-power GPUs.

System simulator implementation. To measure end-to-end accuracy in continuous learning with hardware-based retraining and labeling phases, we develop a system simulator, which uses the accelerator simulator results. It simulates the runtime behaviors of a DCAPO-equipped autonomous system, while the accuracy evaluation is performed on the GPUs. Prior to starting continuous learning, the system generates DNN cycle statistics for the partitioned systolic array, informing the system about the samples processed per phase within a

given timeframe. This information guides the execution of continuous learning jobs on the GPU, aligning with the scheduled hardware allocation. By integrating hardware simulation and GPU kernel execution, our approach simulates end-to-end accuracy and the spatiotemporal allocation algorithm within our continuous learning system.

Baselines. We compare two DCAPO systems with several baseline variants, represented by the combinations of their hardware platforms and resource allocation methods. For both baseline and DCAPO variants, we use two hardware platforms, (1) NVIDIA Orin GPU, and (2) our DCAPO accelerator. We select two state-of-the-art continuous learning systems as the baseline: (1) Ekya [5] sets the retraining configuration through a profiling process at every window, and (2) EOMU [49] selectively triggers the retraining within a shorter window and determines the training configurations by the retraining manager. Below, we describe all the compared continuous learning systems categorized by their combinations of hardware and resource allocation techniques.

- **OrinLow-Ekya.** With this baseline, we examine how power levels influence Ekya’s performance on autonomous systems in low battery conditions. Orin supports multiple power options, ranging from 15W to 60W. “OrinLow” represents Orin using the 30W option, which consumes $127 \times$ more power than DCAPO. This baseline sets the GPU’s maximum frequency to 624.8MHz, the closest to DCAPO’s 500MHz.
- **OrinHigh-Ekya.** We evaluate the performance of Ekya on Orin’s default power setting (i.e., 60W) without power constraints, comparing it to DCAPO on a sufficiently powered autonomous system. This baseline consumes $254 \times$ more power than DCAPO.
- **OrinHigh-EOMU.** We assess the performance of EOMU on Orin GPU at the default power setting. This baseline uses shorter window times to detect data drift and trigger the retraining process. We set 10 seconds as the window time for EOMU according to its paper.
- **DCAPO-Ekya.** This baseline represents a DCAPO-accelerated autonomous system using Ekya’s resource allocation and system configuration mechanism.

B. Accuracy Comparison

Figure 9 presents a comparison of the end-to-end accuracy results among the aforementioned continuous learning system variants. We compare the baselines to the two DCAPO variations with different resource allocation methods: (1) DCAPO-Spatial: optimized static resource allocation within a pre-determined window time, and (2) DCAPO-Spatiotemporal: dynamic spatiotemporal allocation responsive to data drift.

OrinLow-Ekya. When we use Orin with the low-power mode, the entirety of three pairs we used for evaluation offer significantly low accuracy, which ranges from 20.8% to 74.8%. For all experiments, we prioritize resource allocation for inference, incurring no frame drops that largely affect the final accuracy. Therefore, the limited accuracy results is solely attributed to

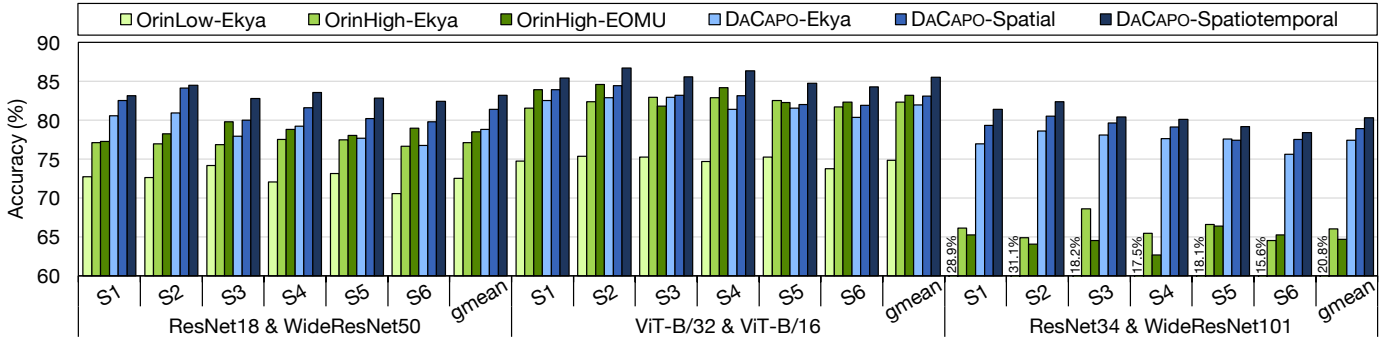


Fig. 9: End-to-end averaged accuracy of different continuously learning autonomous systems on six runtime scenarios.

insufficient computation resources offered for retraining and labeling. In particular, the (ResNet34, WideResNet101) pair demonstrates notably lower accuracy compared to the others since the relatively large model size requires a substantial portion of the Orin GPU’s computational resources for inference, leaving minimal resources available for the other two kernels.

OrinHigh-Ekya and OrinHigh-EOMU. We also compare the Orin GPU baselines with the highest power setting, OrinHigh-Ekya and OrinHigh-EOMU, to examine the full capabilities of NVIDIA’s autonomous system solution. While the two obtain a significant accuracy improvement of respective 11.5% and 13.0% in comparison with OrinLow-Ekya, it is still not closely reaching RTX3090-CL, which suggests potential unleashed benefits of the proposed acceleration solution. Intuitively, when evaluated on the toughest case of (ResNet34, WideResNet101) pair, the two see the largest gains of 45.2% and 43.9% improvements, respectively. We observe a slight difference between Ekya and EMOU since both are based on an assumption that labeling and training are offloaded to remote servers, which makes EOMU’s improved resource allocation method marginally influential on the end accuracy.

DACAPO-accelerated systems. To better understand the implication of the resource allocation algorithm on the DACAPO hardware, we populate three DACAPO-based system variants using our Spatial and Spatiotemporal resource allocation methods along with Ekya. While the DACAPO accelerator offers significantly higher raw performance, we observe that DACAPO-Ekya does not always outperform the Orin baselines significantly. This is because Orin uses significantly higher precision (i.e., 32-bit single-precision FP), while DACAPO leverages low precision (i.e., MX4 to MX9) for higher throughput, which can negatively impact the accuracy behaviors depending on the used models. Particularly for the ViT models, we notice that DACAPO-Ekya performs poorly even compared to the Orin baselines since ViT models often have relatively larger precision sensitivity than other models [61], [98]. Similarly, DACAPO-Spatial offers superior performance compared to the Orin baselines, while the gap is marginal.

On the other hand, when employing DACAPO-Spatiotemporal, the accuracy progresses towards the highest level among all other continuous learning systems, akin to

completing the “last mile” in a journey towards optimal performance under the computational resources of the DACAPO hardware. Concretely, DACAPO-Spatiotemporal achieves significant accuracy gains over OrinHigh-Ekya and OrinHigh-EOMU, which are respectively 6.5% and 5.5% improvements. The results suggest that our hardware-algorithm co-designed acceleration approach effectively unlocks continuous learning’s full capabilities and successfully addresses the “resource wall” challenge.

C. In-Depth Analysis on Resource Allocation Algorithm

We now delve into the in-depth analysis of accuracy change over time in a specific experiment using scenario S1.

Accuracy over time. To better analyze the effectiveness of DACAPO-Spatiotemporal over DACAPO-Spatial and the Orin baselines, Figure 10(a) and Figure 10(b) show accuracy change over an extended timeframe, using (ResNet18, WideResNet50) and (ResNet34, WideResNet101) model pairs. In both pairs, DACAPO-Spatiotemporal works better than DACAPO-Spatial and the Orin baselines as they lack ability to promptly adapt to data drift occurrences. The Ekya and Spatial systems use a fixed predetermined window time, which lacks situational awareness within the dataset. However, DACAPO-Spatiotemporal shows better adaptability to environment changes through iterative retraining as required. EOMU is designed to trigger frequent retrainsings, showing comparatively better accuracy behaviors than Ekya, while the Orin’s resource insufficiency results in its limited gain.

Data drift cases. To analyze the accuracy changes in finer detail, Figure 10(c, d) and Figure 10(e, f) zoom in on specific intervals of interest in Figure 10(a) and Figure 10(b), respectively, allowing for a more focused analysis of accuracy variations observed in these time frames. Figure 10(c) and Figure 10(e) illustrate instances of data drifts where DACAPO-Spatiotemporal demonstrates its effectiveness, swiftly recovering from these drifts via regular retraining sessions. This results in superior performance compared to DACAPO-Spatial, with maximum improvement margins of 13.1% and 12.4% for the two pairs, respectively. Conversely, Figure 10(d) and Figure 10(f) illustrate two suboptimal cases where DACAPO-Spatial outperforms DACAPO-Spatiotemporal. During certain

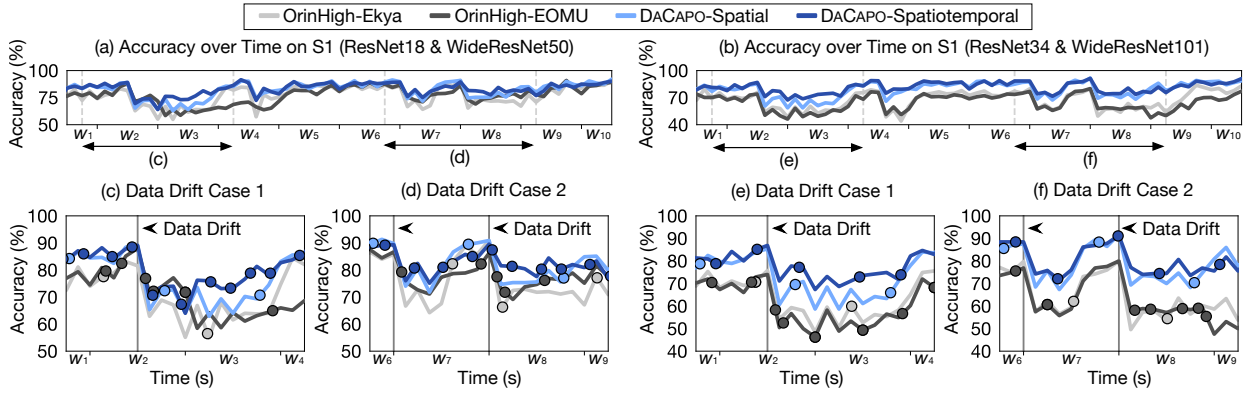


Fig. 10: Accuracy over time. To illustrate the time-series accuracy pattern, we display the average accuracy recorded at 15-second intervals. Each colored dot on the line indicates the completion of a retraining phase.

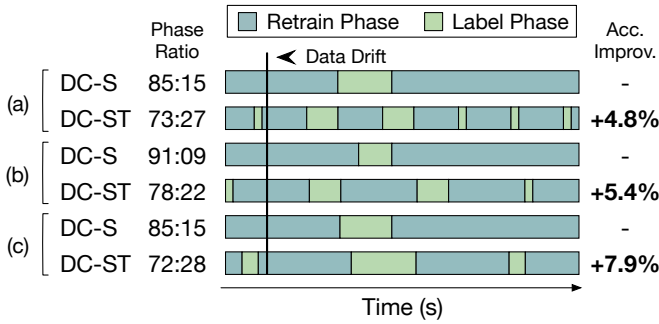


Fig. 11: Scheduling decisions of temporal resource allocation over time: (a) ResNet18 & WideResNet50, (b) ViT-B/32 & ViT-B/16, and (c) ResNet34 & WideResNet101. DC-S is DACAPO-Spatial and DC-ST is DACAPO-Spatiotemporal.

periods, DACAPO-Spatial exhibits higher accuracy, with maximum margins of 5.5% and 5.3%, respectively. However, the overall accuracies of DACAPO-Spatiotemporal and DACAPO-Spatial from w_7 to w_8 stand at 82.4% and 81.7% in Figure 10(d), and 79.7% and 78.0% in Figure 10(f), respectively. Even in its suboptimal cases, DACAPO-Spatiotemporal maintains competitive performance with DACAPO-Spatial.

Additionally, Figure 10(a) and Figure 10(b) also delineate that the Orin baselines almost consistently provide lower accuracy than the DACAPO variants. In fact, the result shows that EOMU triggers substantially more frequent retrains, as marked with the dark grey colored dots. However, such frequent retrains do not always help because trainings with insufficient resource engender incomplete models, which ends up lowering the final accuracy. Overall, the results suggest that DACAPO-Spatiotemporal outperforms all the alternative baselines by leveraging significantly boosted performance of hardware-accelerated computing platform as well as judiciously allocating resources in the spatiotemporal dimension.

D. Effectiveness of Temporal Resource Allocations

Figure 11 shows the results of proposed temporal resource allocations, which are collected during 3 minutes on S1 using

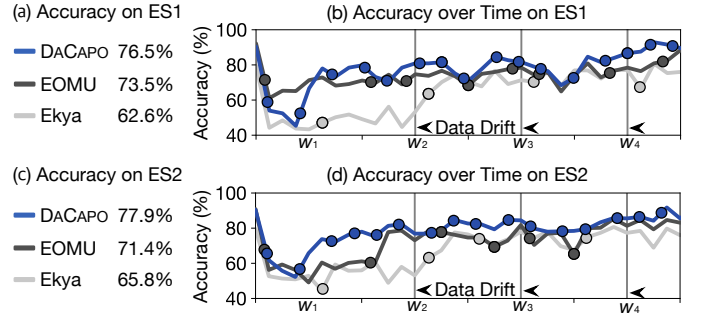


Fig. 12: Accuracy comparison of DACAPO, EOMU, and Ekya, using extreme data drift scenarios. We use (ResNet18 & WideResNet50) model pair for this experiment.

the evaluated three model pairs. For each pair, we present the time breakdown of retraining and labeling phases for DACAPO-Spatial (DC-S) and DACAPO-Spatiotemporal (DC-ST). Additionally, we report the overall breakdown of retraining and labeling on the left and the accuracy improvement results on the right. We observe that when data drifts occur, DACAPO-Spatiotemporal swiftly detects them and allocates 12.7% higher time slots for labeling, achieving 5.9% higher accuracy improvement compared to the spatial-only baseline. These results demonstrate the effectiveness of our strategy that allocates more resources to labeling at the occurrence of data drifts. Note that we achieve the improved accuracy even with “less” retraining time since we cut the assigned time for retraining, exhibiting the effectiveness of rebalancing.

E. Sensitivity Study for Extreme Data Drift Scenarios

Figure 12 compares the accuracy results of three continuous learning systems, when we employ two extreme data drift scenarios (ES1 and ES2) and a model pair (ResNet18 & WideResNet50). In addition to the two extreme scenarios, we evaluate four more extreme scenarios and observe the similar trends, while we only show two cases due to the space limit. The figure reports the averaged accuracies (left) along with the accuracy change over time (right). Compared to the regular

scenario results, Ekya experiences an accuracy degradation of 12.9%. Conversely, EOMU exhibits a higher tolerance than Ekya and offers 7.8% higher accuracy. This improvement is attributed to EOMU’s capability to monitor accuracy changes across short intervals and trigger adaptations by retraining models more frequently than Ekya, as illustrated by dense markers on the line graphs. However, DACAPO, by scheduling spatiotemporal resources for retraining and labeling, achieves the accuracy of 77.2%, which is 4.4% and 13.0% higher than EOMU and Ekya, respectively.

VIII. RELATED WORK

Continuous learning. Continuous learning addresses the critical issue of deployed models becoming outdated over time due to data drift. In the field of machine learning, significant efforts have been directed toward retraining pretrained models with new data without forgetting prior knowledge. The replay-based approach [2], [6], [76], [81], [96] is one such method, integrating valuable historical datasets with new samples for model updates. Alternatively, regularization-based updates [43], [47], [51], [70], [78], [90] augment additional regularization terms during further training to penalize updates on weights critical for previous tasks. However, implementing these solutions often presents practical challenges due to unaddressed training overheads [27], [72]. Distinctively, DACAPO adopts a strategy commonly employed in data streaming scenarios, as exemplified by the video analytics domain.

Live video analytics. Edge devices, positioned near data sources, offer significant benefits, such as bandwidth conservation, reduced latency, and improved privacy by eliminating the need for data transmission to the cloud. However, due to their limited computational capacity, most contemporary edge video analytics systems rely heavily on cloud infrastructure. Prior works [12], [17], [31], [65], [88], [89], [99] aim to use edge devices for preprocessing tasks, such as filtering irrelevant frames or cropping the regions of interest. Several studies [3], [20], [40], [41], [44], [54], [58], [95] propose partitioning DNN computations between edge devices and the cloud. Other works [13], [21], [22], [30], [100] propose enhancing efficiency by caching and reusing previous inference results. Despite these attempts, the involvement of cloud computing constrains the full potential benefits of edge computing, explaining the increasing development of edge-targeted accelerators in the industry [1], [28], [32], [69], [83], [86]. Furthermore, while there are prior works that propose edge-targeted training solutions [11], [14], [45], [53], [62], [64], [94], [102], they do not take into account the tradeoff between resource utilization and final accuracy of training and labeling computations. DACAPO addresses these issues by a tailored hardware accelerator design that enables autonomous systems to perform independent retraining, thus facilitating a novel continuous learning video analytics system.

Low-precision arithmetic for DNN acceleration. Replacing resource-intensive floating point operations with low-precision counterparts is a crucial optimization strategy for DNNs.

Notably, floating point representations with reduced bit width, such as FP16 [33], [63], FP8 [84], [91], BF16 [29], and TF32 [68], introduce certain improvements. However, the inherent computational complexity of floating point often limits their effectiveness. Conversely, fixed-point quantization techniques [15], [39], [57], [75], [103], [104] present computational advantages but often lead to substantial degradation in model accuracy. Block Floating Point (BFP) serves as an effective compromise between these extremes. Courbariaux et al. [16] and Flexpoint [50] introduce approaches to adjust shared exponent values either at fixed intervals or throughout the training process. HybridBFP [23] calculates a shared exponent value for each dot-product operation, enhancing compressibility and accuracy. MSFP [18] underscores the significant role of group sizes and validates their approach on production-level models. FAST [73] utilizes stochastic rounding and a bit-serial hardware architecture to support dynamic bit width adjustment during training. MX [19] introduces additional micro-exponent bits to the shared exponent, showcasing enhanced performance across various domains. DACAPO employs the microexponent concept from MX, effectively fulfilling the need for dynamic precision inference in our novel continuous learning system.

IX. CONCLUSION

Most existing continuous learning systems focus solely on training, neglecting the critical yet compute-heavy tasks of inference and labeling, and typically assume the availability of substantial computation resources, which is often impractical on battery-operated, resource-constrained autonomous systems. To tackle this challenge, this work proposes a hardware-algorithm co-designed solution, DACAPO, which enables concurrent executions of retraining, labeling, and inference in a resource-efficient manner. DACAPO achieves this goal by jointly leveraging a spatially-partitionable, precision-flexible accelerator and a spatiotemporal resource allocation algorithm. Our empirical analyses across six real-world scenarios with varying data drifts demonstrate that DACAPO obtains accuracy improvements of 6.5% and 5.5% over the state-of-the-art systems, Ekya and EOMU, respectively, while consuming $254\times$ less energy. These results confirm that DACAPO is an effective and vital initial step to realize continuous learning on autonomous systems.

ACKNOWLEDGMENTS

We would like to thank Cliff Young, Berkin Akin, and our anonymous reviewers for their review and valuable feedback on the paper. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) (No.2024-00396013, No.2022-0-01037, No.2018-0-00503) under the Graduate School of Artificial Intelligence Semiconductor (IITP-2024-RS-2023-00256472), Information Technology Research Center (ITRC) support program (IITP-2024-2020-0-01795), and Artificial Intelligence Graduate School Program (KAIST) (No.2019-0-00075), funded by the Korea government (MSIT).

REFERENCES

- [1] S. Ahmad, S. Subramanian, V. Boppana, S. Lakka, F.-H. Ho, T. Knopp, J. Noguera, G. Singh, and R. Wittig, "Xilinx first 7nm device: Versal ai core (vc1902)," in *HotChips*, 2019.
- [2] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *CVPR*, 2021.
- [3] G. Bartolomeo, M. Yosofie, S. Baurle, O. Haluszczynski, N. Mohan, and J. Ott, "Oakestra: A lightweight hierarchical orchestration framework for edge computing," in *USENIX ATC*, 2023.
- [4] S. Bateni and C. Liu, "NeuOS: A Latency-Predictable Multi-Dimensional optimization framework for DNN-driven autonomous systems," in *USENIX ATC*, 2020.
- [5] R. Bhardwaj, Z. Xia, G. Ananthanarayanan, J. Jiang, Y. Shu, N. Kariarakis, K. Hsieh, P. Bahl, and I. Stoica, "Ekya: Continuous Learning of Video Analytics Models on Edge Compute Servers," in *NSDI*, 2022.
- [6] Z. Borsos, M. Mutny, and A. Krause, "Coresets via bilevel optimization for continual learning and streaming," *NIPS*, 2020.
- [7] I. Bozcan and E. Kayacan, "AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance," in *ICRA*, 2020.
- [8] I. Bozcan and E. Kayacan, "Context-dependent anomaly detection for low altitude traffic surveillance," in *ICRA*, 2021.
- [9] I. Bozcan, J. Le Fevre, H. X. Pham, and E. Kayacan, "Gridnet: Image-agnostic conditional anomaly detection for indoor surveillance," *IEEE Robotics and Automation Letters*, 2021.
- [10] B. Brown, M. Broth, and E. Vinkhuyzen, "The halting problem: Video analysis of self-driving cars in traffic," in *CHI*, 2023.
- [11] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce memory, not parameters for efficient on-device learning," *Advances in Neural Information Processing Systems*, 2020.
- [12] N. Chen, Y. Chen, Y. You, H. Ling, P. Liang, and R. Zimmermann, "Dynamic urban surveillance video stream processing using fog computing," in *BigMM*, 2016.
- [13] T. Y.-H. Chen, L. Ravindranath, S. Deng, P. Bahl, and H. Balakrishnan, "Glimpse: Continuous, real-time object recognition on mobile devices," in *SenSys*, 2015.
- [14] H.-Y. Chiang, N. Frumkin, F. Liang, and D. Marculescu, "Mobiletl: On-device transfer learning with inverted residual blocks," in *AAAI*, 2023.
- [15] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," *NIPS*, 2015.
- [16] M. Courbariaux, Y. Bengio, and J. David, "Low precision arithmetic for deep learning," in *ICLR*, 2015.
- [17] T. Dao, A. Roy-Chowdhury, N. Nasrabadi, S. V. Krishnamurthy, P. Mohapatra, and L. M. Kaplan, "Accurate and timely situation awareness retrieval from a bandwidth constrained camera network," in *MASS*, 2017.
- [18] B. Darvish Rouhani, D. Lo, R. Zhao, M. Liu, J. Fowers, K. Ovtcharov, A. Vinogradsky, S. Massengill, L. Yang, R. Bittner *et al.*, "Pushing the Limits of Narrow Precision Inference at Cloud Scale with Microsoft Floating Point," in *NeurIPS*, 2020.
- [19] B. Darvish Rouhani, R. Zhao, V. Elango, R. Shafipour, M. Hall, M. Mesmakhosroshahi, A. More, L. Melnick, M. Golub, G. Varatkar *et al.*, "With shared microexponents, a little shifting goes a long way," in *ISCA*, 2023.
- [20] Z. Dong, Q. He, F. Chen, H. Jin, T. Gu, and Y. Yang, "Edgemove: Pipelining device-edge model training for mobile intelligence," in *Proceedings of the ACM Web Conference 2023*, 2023.
- [21] U. Drolia, K. Guo, and P. Narasimhan, "Precog: Prefetching for image recognition applications at the edge," in *SEC*, 2017.
- [22] U. Drolia, K. Guo, J. Tan, R. Gandhi, and P. Narasimhan, "Cachier: Edge-caching for recognition applications," in *ICDCS*, 2017.
- [23] M. Drummond, T. LIN, M. Jaggi, and B. Falsafi, "Training DNNs with Hybrid Block Floating Point," in *NeurIPS*, 2018.
- [24] S. Dutta and C. Ekenna, "Air-to-Ground Surveillance Using Predictive Pursuit," in *ICRA*, 2019.
- [25] S. Fang, Z. Wang, Y. Zhong, J. Ge, S. Chen, and Y. Wang, "TBP-Former: Learning Temporal Bird's-Eye-View Pyramid for Joint Perception and Prediction in Vision-Centric Autonomous Driving," in *CVPR*, 2023.
- [26] S. Ghodrati, B. H. Ahn, J. K. Kim, S. Kinzer, B. R. Yatham, N. Alla, H. Sharma, M. Alian, E. Ebrahimi, N. S. Kim *et al.*, "Planaria: Dynamic architecture fission for spatial multi-tenant acceleration of deep neural networks," in *MICRO*, 2020.
- [27] Y. Ghunaim, A. Bibi, K. Alhamoud, M. Alfarra, H. A. A. K. Hammoud, A. Prabhu, P. H. Torr, and B. Ghanem, "Real-time evaluation in online continual learning: A new hope," in *CVPR*, 2023.
- [28] Google, "Edge tpu," <https://cloud.google.com/edge-tpu/>, 2018.
- [29] Google, "Bfloat16: The secret to high performance on cloud tpus," <https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus?hl=en>, 2019.
- [30] P. Guo, B. Hu, R. Li, and W. Hu, "Foggycache: Cross-device approximate computation reuse," in *MobiCom*, 2018.
- [31] Y. Guo, B. Zou, J. Ren, Q. Liu, D. Zhang, and Y. Zhang, "Distributed and efficient object detection via interactions among devices, edge, and cloud," *ToMM*, 2019.
- [32] Hailo, "Hailo expands hail-8 ai accelerator portfolio," <https://hailo.ai/hailo-8l-entry-level-ai-accelerator-announcement/>, 2023.
- [33] N.-M. Ho and W.-F. Wong, "Exploiting half precision arithmetic in nvidia gpus," in *HPEC*, 2017.
- [34] J. Houyon, A. Cioppa, Y. Ghunaim, M. Alfarra, A. Halin, M. Henry, B. Ghanem, and M. V. Droogenbroeck, "Online Distillation With Continual Learning for Cyclic Domain Shifts," in *CVPR*, 2023.
- [35] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning," in *ECCV*, 2022.
- [36] S. Huai, D. Liu, H. Kong, X. Luo, W. Liu, R. Subramaniam, C. Makaya, and Q. Lin, "Collate: Collaborative Neural Network Learning for Latency-Critical Edge Systems," in *ICCD*, 2022.
- [37] S. Huai, L. Zhang, D. Liu, W. Liu, and R. Subramaniam, "ZeroBN: Learning Compact Neural Networks For Latency-Critical Edge Systems," in *DAC*, 2021.
- [38] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," in *ICCV*, 2023.
- [39] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," *NIPS*, 2016.
- [40] C.-C. Hung, G. Ananthanarayanan, P. Bodik, L. Golubchik, M. Yu, P. Bahl, and M. Philipose, "Videoeage: Processing camera streams using hierarchical clusters," in *SEC*, 2018.
- [41] L. Jia, Z. Zhou, F. Xu, and H. Jin, "Cost-efficient continuous edge learning for artificial intelligence of things," *IEEE Internet of Things Journal*, 2022.
- [42] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *CVPR*, 2023.
- [43] S. Jung, H. Ahn, S. Cha, and T. Moon, "Continual learning with node-importance based adaptive group sparse regularization," *NIPS*, 2020.
- [44] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," in *ASPLOS*, 2017.
- [45] O. Khan, G. Park, and E. Seo, "Dacapo: An on-device learning scheme for memory-constrained embedded systems," *ACM Transactions on Embedded Computing Systems*, 2023.
- [46] M. Khani, G. Ananthanarayanan, K. Hsieh, J. Jiang, R. Netravali, Y. Shu, M. Alizadeh, and V. Bahl, "RECL: Responsive Resource-Efficient Continuous Learning for Video Analytics," in *NSDI*, 2023.
- [47] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *PNAS*, 2017.
- [48] H. Kong, D. Liu, X. Luo, S. Huai, R. Subramaniam, C. Makaya, Q. Lin, and W. Liu, "Towards Efficient Convolutional Neural Network for Embedded Hardware via Multi-Dimensional Pruning," in *DAC*, 2023.
- [49] Y. Kong, P. Yang, and Y. Cheng, "Edge-Assisted On-Device Model Update for Video Analytics in Adverse Environments," in *MM*, 2023.
- [50] U. Köster, T. Webb, X. Wang, M. Nassar, A. K. Bansal, W. Constable, O. Elibol, S. Gray, S. Hall, L. Hornof *et al.*, "Flexpoint: An Adaptive Numerical Format for Efficient Training of Deep Neural Networks," in *NeurIPS*, 2017.
- [51] J. Lee, H. G. Hong, D. Joo, and J. Kim, "Continual learning with extended kronecker-factored approximate curvature," in *CVPR*, 2020.
- [52] J. Lee, J. Choi, J. Kim, J. Lee, and Y. Kim, "Dataflow mirroring: Architectural support for highly efficient fine-grained spatial multitasking on systolic-array npus," in *DAC*, 2021.

- [53] S. Lee, M. Weerakoon, J. Choi, M. Zhang, D. Wang, and M. Jeon, "CarM: Hierarchical episodic memory for continual learning," in *DAC*, 2022.
- [54] D. Li, T. Salonidis, N. V. Desai, and M. C. Chuah, "Deepcham: Collaborative edge-mediated adaptive deep learning for mobile object recognition," in *SEC*, 2016.
- [55] S. Li, Y. Yang, D. Zeng, and X. Wang, "Adaptive and background-aware vision transformer for real-time uav tracking," in *ICCV*, 2023.
- [56] Y. Li, D. Yuan, M. Sun, H. Wang, X. Liu, and J. Liu, "Generalized UAV Object Detection via Frequency Domain Disentanglement," in *CVPR*, 2023.
- [57] D. Lin, S. Talathi, and S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *ICML*. PMLR, 2016, pp. 2849–2858.
- [58] P. Liu, B. Qi, and S. Banerjee, "Edgeeye: An edge service framework for real-time intelligent video analytics," in *EdgeSys*, 2018.
- [59] S. Liu, X. Li, H. Lu, and Y. He, "Multi-Object Tracking Meets Moving UAV," in *CVPR*, 2022.
- [60] S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, and Q. Wu, "AerialVn: Vision-and-language navigation for uavs," in *ICCV*, 2023.
- [61] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, "Post-Training Quantization for Vision Transformer," in *NeurIPS*, 2021.
- [62] X. Ma, S. Jeong, M. Zhang, D. Wang, J. Choi, and M. Jeon, "Cost-effective on-device continual learning over memory hierarchy with miro," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023.
- [63] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed Precision Training," in *ICLR*, 2018.
- [64] R. Miles, M. K. Yucel, B. Manganello, and A. Saà-Garriga, "Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation," in *CVPR*, 2023.
- [65] C. S. Mishra, J. Sampson, M. T. Kandemir, V. Narayanan, and C. R. Das, "Usas: A sustainable continuous-learning framework for edge servers," in *HPCA*, 2024.
- [66] R. T. Mullapudi, S. Chen, K. Zhang, D. Ramanan, and K. Fatahalian, "Online Model Distillation for Efficient Video Inference," in *ICCV*, 2019.
- [67] S.-H. Noh, J. Koo, S. Lee, J. Park, and J. Kung, "FlexBlock: A Flexible DNN Training Accelerator with Multi-Mode Block Floating Point Support," in *arXiv*, 2022.
- [68] NVIDIA, "Getting immediate speedups with nvidia a100 tf32," <https://developer.nvidia.com/blog/getting-immediate-speedups-with-a100-tf32/>, 2020.
- [69] NVIDIA, "Nvidia jetson," <https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/>, 2023.
- [70] I. Paik, S. Oh, T. Kwak, and I. Kim, "Overcoming catastrophic forgetting by neuron-level plasticity control," in *AAAI*, 2020.
- [71] L. Pichierrri, G. Carnevale, L. Sforni, A. Testa, and G. Notarstefano, "A distributed online optimization strategy for cooperative robotic surveillance," in *ICRA*, 2023.
- [72] A. Prabhu, H. A. A. K. Hammoud, P. Dokania, P. H. Torr, S.-N. Lim, B. Ghanem, and A. Bibi, "Computationally budgeted continual learning: What does matter?" in *CVPR*, 2023.
- [73] S. Qian Zhang, B. McDanel, and H. T. Kung, "FAST: DNN Training Under Variable Precision Block Floating Point with Stochastic Rounding," in *HPCA*, 2022.
- [74] E. Qin, A. Samajdar, H. Kwon, V. Nadella, S. Srinivasan, D. Das, B. Kaul, and T. Krishna, "SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training," in *HPCA*, 2020.
- [75] R. Razani, G. Morin, E. Sari, and V. P. Nia, "Adaptive binary-ternary quantization," in *CVPR*, 2021.
- [76] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017.
- [77] A. Samajdar, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "Scale-sim: Systolic cnn accelerator simulator," *arXiv preprint arXiv:1811.02883*, 2018.
- [78] S. S. Sarwar, A. Ankit, and K. Roy, "Incremental learning in deep convolutional neural networks using partial network sharing," *PAML*, 2019.
- [79] M. Seo, D. Cho, S. Lee, J. Park, D. Kim, J. Lee, J. Ju, H. Noh, and D.-G. Choi, "A self-supervised sampler for efficient action recognition: Real-world applications in surveillance systems," *IEEE Robotics and Automation Letters*, 2022.
- [80] H. Sharma, J. Park, N. Suda, L. Lai, B. Chau, V. Chandra, and H. Esmaeilzadeh, "Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural networks," in *ISCA*, 2018.
- [81] D. Shim, Z. Mai, J. Jeong, S. Sanner, H. Kim, and J. Jang, "Online class-incremental continual learning with adversarial shapley value," in *AAAI*, 2021.
- [82] M. K. Shirkoohi, P. Hamadani, A. Nasr-Esfahany, and M. Alizadeh, "Real-Time Video Inference on Edge Devices via Adaptive Model Streaming," in *CVPR*, 2021.
- [83] A. Skillman and T. Edso, "A technical overview of cortex-m55 and ethos-u55: Arm's most capable processors for endpoint ai," in *HotChips*, 2020.
- [84] X. Sun, J. Choi, C.-Y. Chen, N. Wang, S. Venkataramani, V. V. Srinivasan, X. Cui, W. Zhang, and K. Gopalakrishnan, "Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks," *NIPS*, 2019.
- [85] A. Suprem, J. Arulraj, C. Pu, and J. E. Ferreira, "ODIN: Automated Drift Detection and Recovery in Video Analytics," in *VLDB*, 2020.
- [86] E. Talpes, D. D. Sarma, G. Venkataraman, P. Bannan, B. McGee, B. Floering, A. Jalote, C. Hsiong, S. Arora, A. Gorti *et al.*, "Compute solution for tesla's full self-driving computer," *MICRO*, 2020.
- [87] N. Tijtgat, W. Van Ranst, T. Goedeme, B. Volckaert, and F. De Turck, "Embedded Real-Time Object Detection for a UAV Warning System," in *ICCV*, 2017.
- [88] J. Wang, Z. Feng, Z. Chen, S. George, M. Bala, P. Pillai, S.-W. Yang, and M. Satyanarayanan, "Bandwidth-efficient live video analytics for drones via edge computing," in *SEC*, 2018.
- [89] L. Wang, K. Lu, N. Zhang, X. Qu, J. Wang, J. Wan, G. Li, and J. Xiao, "Shoggoth: Towards efficient edge-cloud collaborative real-time video inference via adaptive online learning," in *DAC*, 2023.
- [90] L. Wang, M. Zhang, Z. Jia, Q. Li, C. Bao, K. Ma, J. Zhu, and Y. Zhong, "Afec: Active forgetting of negative transfer in continual learning," *NIPS*, 2021.
- [91] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan, "Training deep neural networks with 8-bit floating point numbers," *NIPS*, 2018.
- [92] T. Wang, G. Dong, and P. Huang, "Pheromone-diffusion-based conscientious reactive path planning for road network persistent surveillance," in *ICRA*, 2021.
- [93] X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Neural map prior for autonomous driving," in *CVPR*, 2023.
- [94] L. Yang, A. S. Rakin, and D. Fan, "Repnet: Efficient on-device learning via feature reprogramming," in *CVPR*, 2022.
- [95] S. Yi, Z. Hao, Q. Zhang, Q. Zhang, W. Shi, and Q. Li, "Lavea: Latency-aware video analytics on edge computing platform," in *SEC*, 2017.
- [96] J. Yoon, D. Madaan, E. Yang, and S. J. Hwang, "Online coreset selection for rehearsal-based continual learning," in *ICLR*, 2022.
- [97] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
- [98] Z. Yuan, C. Xue, Y. Chen, Q. Wu, and G. Sun, "PTQ4ViT: Post-training Quantization for Vision Transformers with Twin Uniform Quantization," in *ECCV*, 2022.
- [99] C. Zhang, Q. Cao, H. Jiang, W. Zhang, J. Li, and J. Yao, "Ffs-va: A fast filtering system for large-scale video analytics," in *ICPP*, 2018.
- [100] L. Zhang, G. Gao, and H. Zhang, "Towards Data-Efficient Continuous Learning for Edge Video Analytics via Smart Caching," in *SenSys*, 2022.
- [101] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline," in *CVPR*, 2022.
- [102] S. Q. Zhang, T. Tambe, N. Cuevas, G.-Y. Wei, and D. Brooks, "Camel: Co-designing ai models and edrams for efficient on-device learning," in *HPCA*, 2024.
- [103] S. Zhou, Z. Ni, X. Zhou, H. Wen, Y. Wu, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *CoRR*, 2016.
- [104] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," in *ICLR*, 2017.