

Enabling On-Device Continual Learning with Binary Neural Networks and Latent Replay

Lorenzo Vorabbi^{1,2}^a, Davide Maltoni²^b, Guido Borghi²^c and Stefano Santi¹

¹Datalogic Labs, Bologna, 40012, Italy

²Department of Computer Science and Engineering (DISI), University of Bologna, Cesena, Italy

Keywords: Binary Neural Networks, On-Device Learning, TinyML, Continual Learning.


Abstract: On-device learning remains a formidable challenge, especially when dealing with resource-constrained devices that have limited computational capabilities. This challenge is primarily rooted in two key issues: first, the memory available on embedded devices is typically insufficient to accommodate the memory-intensive back-propagation algorithm, which often relies on floating-point precision. Second, the development of learning algorithms on models with extreme quantization levels, such as Binary Neural Networks (BNNs), is critical due to the drastic reduction in bit representation. In this study, we propose a solution that combines recent advancements in the field of Continual Learning (CL) and Binary Neural Networks to enable on-device training while maintaining competitive performance. Specifically, our approach leverages binary latent replay (LR) activations and a novel quantization scheme that significantly reduces the number of bits required for gradient computation. The experimental validation demonstrates a significant accuracy improvement in combination with a noticeable reduction in memory requirement, confirming the suitability of our approach in expanding the practical applications of deep learning in real-world scenarios.


1 INTRODUCTION


In recent times, the integration of Artificial Intelligence into the Internet of Things (IoT) paradigm (Mohamed, 2020; Alshehri and Muhammad, 2020), enabling the provision of intelligent systems capable of learning even within embedded or tiny devices, has garnered significant attention in the literature. This trend has been facilitated by various factors, including the evolution of microchips, which have led to the availability of cost-effective chips in many everyday objects. Additionally, the exploration of new learning paradigms, such as Continual Learning (CL) (Parisi et al., 2019; Masana et al., 2022), has contributed to the development of techniques for training neural networks continuously, on small data portions (denoted as *experiences*) at a time, mitigating the issue of catastrophic forgetting (Kirkpatrick et al., 2017). In this manner, a neural network, in contrast to the traditional machine learning paradigm, does not learn from a single large dataset accessible entirely during the train-

ing phase but rather from small data portions accessible gradually over time. This limited amount of data needed by the training procedure effectively simplifies the adoption of a CL training implementation on embedded devices.

Despite the keen interest of the scientific community, numerous challenges still persist, rendering the utilization of deep learning models on devices particularly demanding. These challenges are primarily associated with the computational requirements typically demanded by deep neural networks, even though based on CL strategies. Indeed, embedded devices often have limited available memory, preventing the storage of a vast amount of data. Furthermore, a powerful GPU is usually absent due to cost, space constraints, and energy consumption. These competing needs have given rise in the last few years to a specific branch of machine learning and deep learning called TinyML (Banbury et al., 2020), focused on shrinking and compressing neural network models with respect to the target device characteristics. One of the most interesting TinyML approaches, is Binary Neural Networks (BNNs) (Courbariaux et al., 2016; Rastegari et al., 2016; Qin et al., 2020), where a single bit is used to encode weights and activations; un-

^a <https://orcid.org/0000-0002-4634-2044>

^b <https://orcid.org/0000-0002-6329-6756>

^c <https://orcid.org/0000-0003-2441-7524>

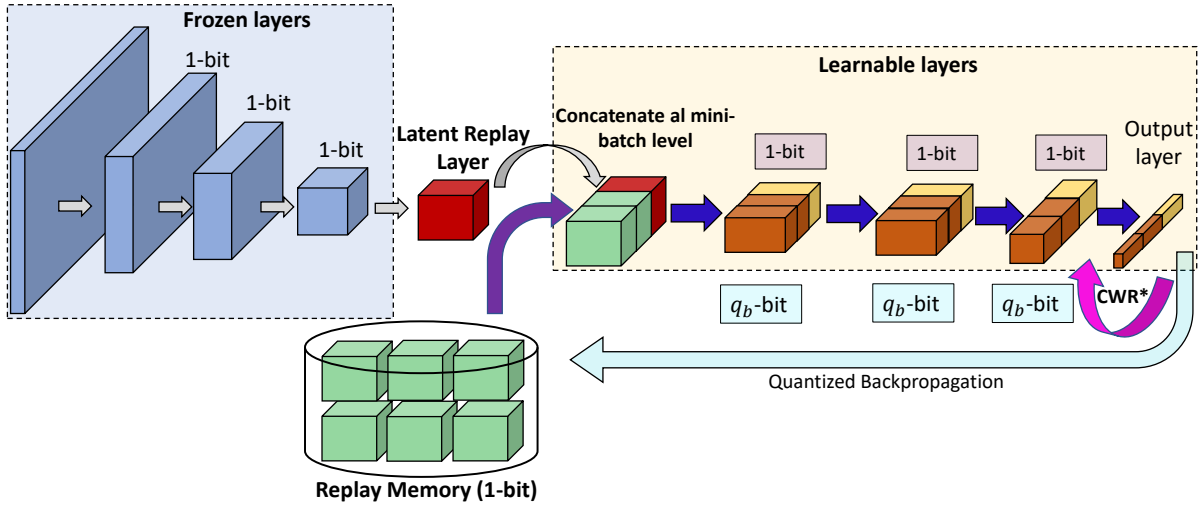


Figure 1: Continual Learning with latent replay memory. When using a BNN the activations stored in the replay memory can be quantized to 1-bit.

fortunately, solutions based on BNNs in combination with Continual Learning algorithms are still lacking.

A previous work (Vorabbi et al., 2023a) explored the possibility of training a BNN model on-device by freezing the binary backbone and allowing the adaptation of only the last classification layer where forgetting is mitigated by CWR* (Lomonaco et al., 2020; Graffieti et al., 2022). Unfortunately, the reported results are interesting but the final accuracy is significantly lower w.r.t. a system where all the layers can be tuned. Pellegrini et al. (Pellegrini et al., 2020) showed that a good accuracy/efficiency tradeoff in CL can be achieved by only some convolutional layers (typically from 3 to 5), placed before the classification head. Replaying part of old data (stored in a replay memory or buffer), interleaved with new samples, was proved to be an effective approach to mitigate catastrophic forgetting (Kirkpatrick et al., 2017). If past samples are stored as intermediate activations (instead of raw data), the replay technique takes the name *latent replay* (Pellegrini et al., 2020) (see Fig. 1). Latent replay is particularly interesting when combined with BNN (as proposed in this paper) since the latent activations can be quantized to 1-bit, leading to a remarkable storage saving. Unfreezing some intermediate layers requires to back-propagate gradients along the model to update weights; on the edge, the implementation of this process, usually referred to as *on-device learning*, requires an efficient and lightweight back-propagation implementation, which is not yet available in the most popular training frameworks. The reduction of bitwidths during backward pass, made possible by a fixed point (many low-power CPUs are not equipped with floating-point unit) implementa-

tion, can speedup the learning phase but the tradeoffs between accuracy loss and efficiency need to be evaluated with attention.

In this paper, we propose a solution to combine the Continual Learning paradigm with training on the edge using BNNs. Specifically, through the introduction of a back-propagation and input binarization algorithm, we demonstrate how it is possible to continuously tune a CNN model (including classification head and convolutional layers) with low memory requirements and high efficiency. Our work represents a step beyond the classical quantization approach of BNNs published in the literature, where binarization is typically considered only during forward pass and a binary model is trained using latent floating-point weights (Helweg et al., 2019). Some works showed (Cai et al., 2020; Lin et al., 2022) good improvements in reducing both the memory demand and the computational effort to enable training on the edge, but they did not focus on the Continual Learning (CL) scenario, which we primarily address. We conducted experiments with multiple BNN models, evaluating the advantages offered by the proposed methodology in comparison to the method outlined in (Vorabbi et al., 2023a) where only the classification head is tuned.

The main contributions of this work can be summarized as follows:

1. **Reduced Replay Memory Requirement:** our replay memory stores intermediate activations quantized to 1-bit allowing a relevant storage saving. We investigate the trade-offs required to maintain model accuracy while simultaneously reducing memory consumption.

2. **Improved Model Accuracy:** by enabling the continual adaptation of intermediate convolutional layers (besides the final classification head) our BNN-based model significantly outperforms the closest previous solution (Vorabbi et al., 2023a).
3. **Quantization of Backpropagation for Non-Binary Layers:** we introduce a quantization approach for the back-propagation step in non-binary layers, enabling the preservation of accuracy while eliminating floating-point operations.
4. **Optimized Binary Weight Quantization:** we present an optimized quantization strategy tailored for binary weights, leading to a remarkable $8\times$ reduction in memory requirements. Binary layers are typically trained by storing latent floating-point representations of weights that are subsequently binarized during inference. Replicating this schema on-device would result in an unacceptable increase of memory usage and computational overhead.
5. **Optimized Back-Propagation Framework:** we implemented a comprehensive back-propagation framework capable of supporting various quantization levels both inference and back-propagation stages.

The paper is organized as follows. In Section 3.1 we describe the latent replay mechanism providing an estimation of the memory saved when applied to a binary layer. In Section 3.2 we detail the quantization approach used for both forward and backward passes. Then, in Section 3.3 we describe the method used to quantize gradient computation. A comprehensive experimental evaluation is proposed in Section 4, focusing on the accuracy comparison with respect to the CWR* algorithm (Sect. 4.1), the reduction of the storage needed by the replay memory (Sect. 4.2) and the efficiency in the backpropagation algorithm (Sect. 4.4).

2 RELATED WORK

Continual Learning: Some works in the literature addressed the on-device learning task proposing solutions to primarily reduce the memory requirement of the learning algorithm: Ren et al. (Ren et al., 2021) brought the transfer learning task on tiny devices by adding a trainable layer on top of a frozen inference model. Cai et al. (Cai et al., 2020) proposed to freeze the model weights and retrain only the biases reducing the memory storage during forward pass. Lin et al. (Lin et al.,

2022) introduced a sparse update technique to skip the gradient computation of less important layers and sub-tensors. QLR-CL (Ravaglia et al., 2021) relies on low-bitwidth quantization (8-bit) to speed up the execution of the network up to the latent layer and at the same time reduce the memory requirement of the latent replay vectors from the 32-bit floating point to 8-bit; compared to our solution, QLR-CL optimizes the computation pipeline for a specific ultra low-power CPU based on RISC-V ISA. In addition, backpropagation is performed with floating-point precision. In (Nadalini et al., 2022; Nadalini et al., 2023), Nadalini et al. introduced a framework to execute on-device learning on tiny devices using floating-point (32 and 16 bits) computation. Our solution differs considerably even in this case because we introduce a quantized fixed-point implementation for binary and non-binary layers instead of performing a post-training quantization of the frozen layers and then executing backpropagation with floating-point precision. Additionally, to the best of the author’s knowledge, this is the first work that implements on-device learning by quantizing the back-prop of binary layers using low bitwidths.

As in (Vorabbi et al., 2023a) the proposed approach uses CWR* for class-bias correction in the classification head (see (Masana et al., 2022)). CWR* maintains two sets of weights for the output classification layer: cw are the consolidated weights used during inference while tw are the temporary weights that are iteratively updated during back-propagation. cw are initialized to 0 before the first batch and then updated according to Algorithm 1 of (Lomonaco et al., 2020), while tw are reset to 0 before each training mini-batch. CWR*, for each already encountered class (of the current training batch), reloads the consolidated weights cw at the beginning of each training batch and, during the consolidation step, adopts a weighted sum based on the number of the training samples encountered in the past batches and those of current batch.

Binary Neural Networks: Quantization is a useful technique to compress Neural Network models compared to their floating-point counterparts, by representing the network weights and activations with very low precision. The most extreme quantization is binarization, where data can only have two possible values, namely -1 (0) or $+1$ (1). By representing weights and activations using only 1-bit, the resulting memory footprint of the model is dramatically reduced and the heavy matrix mul-

tiplication operations can be replaced with light-weighted bitwise XNOR operations and Bitcount operations. According to (Bannink et al., 2021), which compared the speedups of binary layers w.r.t. the 8-bit quantized and floating point layers, a binary implementation can achieve a lower inference time from 9 to 12 \times on a low power ARM CPU. Therefore, Binary Neural Networks combine many hardware-friendly properties including memory saving, power efficiency and significant acceleration; for some network topologies, BNN can be executed on-device without the usage of floating-point operations (Vorabbi et al., 2023b) simplifying the deployment on ASIC or FPGA hardware.

3 METHOD

In this section, we introduce our solution to efficiently deploy CL methods using Latent Replay and BNNs. In particular, the CWR* approach (briefly summarized in Sect. 2) is used to correct class-bias in the classification head.

3.1 Continual Learning with Latent Replays

In Fig. 1 we illustrate the CL process with Latent Replay. When new data becomes available, they are fed to the neural network that during the forward pass produces their latent activations, which represent the feature maps corresponding to a specific intermediate layer. We denote this layer as l (where $l \in [0, L)$), with L representing the total number of layers within the model. Activations of new data are joined (at minibatch level) with the replay activations (previously stored) and forward/backward passes on the remaining layers, specifically those with index from $l + 1$ to $L - 1$. To elucidate further, if B_N denotes the minibatch size of the newly acquired latent activations, a subset of replay vectors (B_R) is extracted from the replay memory and merged, thus forming a minibatch of total size $B_T = B_N + B_R$. In contrast, the layers with an index less than l are maintained in a frozen state and are not included in the learning process. After the conclusion of each training experience, the replay memory is updated by including samples from the last experience and using class-balanced reservoir sampling (Vitter, 1985), which ensures a double balancing: (i) in terms of samples per classes, (ii) in terms of samples from experience (see Algorithm 1).

Algorithm 1: Procedure used to populate the replay memory (RM). RM is initially pre-populated using training samples of the first experience. The reservoir sampling is used on a class basis to maintain the balance among different classes. This approach prevents a skewed representation of classes within RM.

Input: $N = \max$ number of samples per class
Input: $C = \max$ number of classes

- 1 $RM_{size} = C \cdot N$;
// $C \cdot N$ is the max size of RM populated during the first experience.
- 2 **for** each on-device experience **do**
// T is the number of classes
// $M_t =$ samples of class t
// $RM_t =$ samples of class t already in RM
- 3 **for** $t = 0$ to $T - 1$ **do**
- 4 $B_t = RM_t \cup M_t$;
// # is the cardinality operator
- 5 $RM_t^{new} =$ apply Reservoir sampling to extract # RM_t samples from B_t ;
- 6 remove not selected RM_t samples ;
- 7 update RM with RM_t^{new} ;
- 8 **end**
- 9 **end**

3.2 Quantization of Activations and Weights

Quantization techniques have gained widespread adoption to diminish the data size associated with model parameters and the activations of layers. Employing quantization strategies enables the reduction of data bitwidth from the conventional 32-bit floating-point representation to a lower bit-precision format, typically 8 bits or even less, while typically incurring a negligible loss in accuracy during the forward pass of the model. For the quantization of non-binary layers that need to be trained on-device, we adopted the approach proposed in the work of Jacob et al. (Jacob et al., 2018) which is further implemented in the GEMMLOWP library (Jacob et al., 2017).

By representing the dynamic range of the activations at the i -th layer of the network as $[a_{min}^i, a_{max}^i]$, we can define the quantized activations a_q as:

$$a_q^i = \text{cast}_{10,q} \lfloor \frac{a^i}{S_a^i} \rceil, \quad S_a^i = \frac{a_{max}^i - a_{min}^i}{2^q - 1} \quad (1)$$

where q denotes the number of quantization bits used (8, 16, 32), a^i represents the full-precision activations and a_{max}^i, a_{min}^i are determined through calibration on

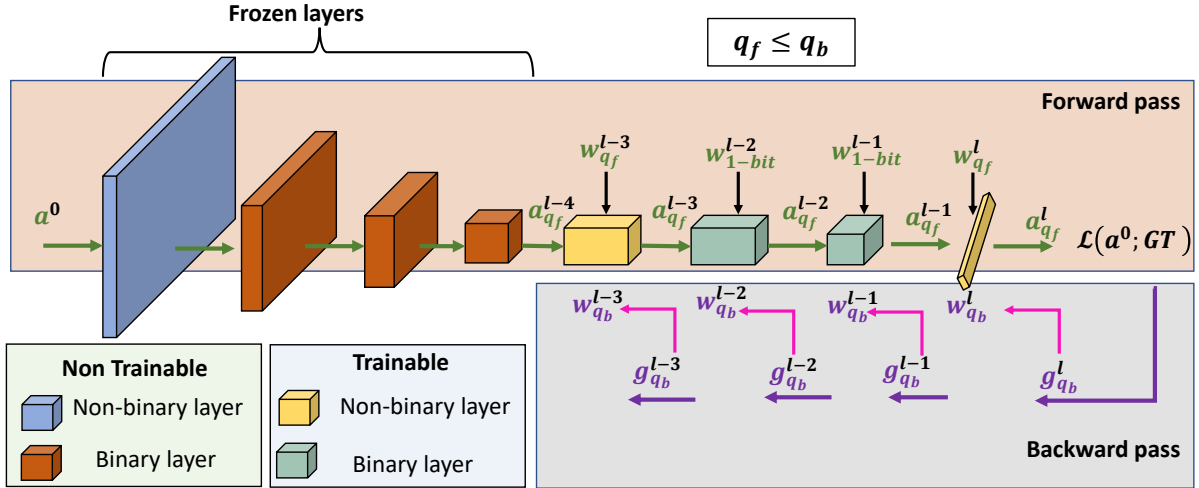


Figure 2: Quantization scheme that uses a different number of bitwidth for forward (q_f) and backward (q_b) pass. Usually, trainable non-binary layers are Batch Normalization (Ioffe and Szegedy, 2015), Addition and Concatenation layers.

the training dataset. Weight quantization can be accomplished using an equation analogous to equation 1. However, as recommended in Vorabbi et al. (Vorabbi et al., 2023a), we utilize two separate sets of quantization bits for both the forward and backward passes. For binary layers, during the forward pass, binarization is executed according to the following equation:

$$STE(x) = \begin{cases} +1 & x \geq 0 \\ -1 & otherwise \end{cases} \quad (2)$$

as proposed in (Courbariaux et al., 2016). In backward pass, STE computes the derivative of sign as if the binary operation was a linear function. This approximation has been further improved by other works (Liu et al., 2018; Liu et al., 2020) and in general it is model dependent.

3.3 Quantized Backpropagation

Drawing upon the findings presented in the works of Gupta et al. (Gupta et al., 2015), Das et al. (Das et al., 2018), and Banner et al. (Banner et al., 2018), it is evident that the quantization of gradients stands out as the primary contributor to accuracy degradation during the training process. Therefore, we advocate for a quantization scheme akin to that introduced in (Vorabbi et al., 2023a). In this scheme, we employ two distinct sets of quantization bits for the forward and backward passes.

The back-propagation algorithm operates in an iterative manner to calculate the gradients of the loss function (denoted as \mathcal{L}) with respect to the input a^{l-1} for the layer l :

$$g^l = \frac{\partial \mathcal{L}}{\partial a^l} \quad (3)$$

starting from the last layer. Every layer in the network is tasked with computing two sets of gradients to execute the iterative update process. The first set corresponds to the layer activation gradient w.r.t. the inputs, which serves the purpose of propagating gradients backward to the previous layer. Considering a linear layer, where $a^l = W^l \cdot a^{l-1}$ and $\frac{\partial a^l}{\partial a^{l-1}} = W^l$, the gradients can be computed as follows:

$$g^{l-1} = \frac{\partial \mathcal{L}}{\partial a^l} \cdot \frac{\partial a^l}{\partial a^{l-1}} = W^l g^l \quad (4)$$

The other set is used to update the weights of layer index l :

$$g_w^l = \frac{\partial \mathcal{L}}{\partial a^l} \cdot \frac{\partial a^l}{\partial W^l} = a^{l-1} g^l \quad (5)$$

Based on Eq. 4 and 5, the backward pass requires approximately twice Multiply-And-Accumulate (MAC) operations compared to the forward pass and therefore the gradient quantization becomes essential to efficiently train neural network models on-device. The quantization of weights and gradients (Eq. 4 and 5) is implemented through Eq. 1 and can be visually summarized in Fig. 2; as shown in (Banner et al., 2018; Vorabbi et al., 2023a), backward pass usually needs higher bitwidth to preserve the directionality of the weight tensor and, based on that, we propose to use lower bitwidth during forward pass (Fig. 2, q_f bits, green path) to minimize latency and more bits for the backward pass to be more accurate in gradient representation (Fig. 2, q_b bits, purple path). Considering the constrained memory resources available on embedded devices, accurately estimating

Table 1: The table represents a comparison of memory usage (# parameters) for different BNN models. With B we report the number of binary weights that can be updated during back-propagation; with NB the number of non-binary weights. The choice of latent replay (LR) level is discussed in Section 4. It is worth noting that the largest part of memory weights is used by binary weights.

	# total weights	LR shape	# B = binary weights after LR	# NB = non-binary weights after LR	$\frac{B}{B+NB}$
<i>BiReal-18</i>	11.2M	(4, 4, 512)	7.0M	19K	99.7%
<i>BiReal-18</i>	11.2M	(8, 8, 256)	10.1M	28K	99.7%
<i>React-18</i>	11.1M	(4, 4, 512)	7.0M	18K	99.7%
<i>React-18</i>	11.1M	(8, 8, 256)	8.3M	24K	99.7%
<i>VGG-Small</i>	4.6M	(8, 8, 512)	2.3M	86K	96.4%
<i>QuickNet</i>	12.7M	(7, 7, 256)	9.5M	36K	99.6%
<i>QuickNet</i>	12.7M	(14, 14, 128)	11.9M	43K	99.6%
<i>QuickNetLarge</i>	22.8M	(7, 7, 256)	14.2M	40K	99.7%
<i>QuickNetLarge</i>	22.8M	(14, 14, 128)	21.3M	56K	99.7%

the memory requirements of the learning algorithm becomes imperative. We can categorize memory into two distinct types: the memory utilized by the CL method (e.g. the replay memory) and the memory necessary to store intermediate tensors during the forward pass, which are subsequently used in the back-propagation, along with the model weights. In this context, we will focus mainly on the latter aspect, particularly for binary layers where q_f is fixed at 1-bit while q_b can vary depending on the desired level of accuracy. In Table 1, we present an assessment of the memory usage for representing binary weights of trainable layers on-device. It is worth noting that binary weights, as indicated in the fifth column of the same table, constitute a substantial portion of the total model parameters. Consequently, reducing q_b to 1-bit offers significant memory savings in comparison to a more conventional approach where q_b is set to 16 bits. The reduction in memory usage exhibits an almost linear relationship with the number of bits utilized. We distinguish q_b between binary and non-binary layers to apply different quantization bitwidths, as elaborated in Section 4, which demonstrates that it is feasible to maintain accuracy while significantly reducing q_b for binary layers. Denoting q_b^{bin} and $q_b^{non-bin}$ as the quantization settings for binary and non-binary layers, respectively, in Section 4 we illustrate that setting q_b^{bin} to 1-bit results in minimal accuracy loss compared to higher quantization bitwidths.

4 EXPERIMENTS

We evaluate our methods on three classification datasets: CORE50(Lomonaco and Maltoni, 2017), CIFAR10 (Krizhevsky et al., 2009) and CIFAR100(Krizhevsky et al., 2009) with different

BNN architectures. The BNN models employed for CORE50 have been pre-trained on ImageNet through the Larq repository¹; differently, the models used for CIFAR10 and CIFAR100 have been pre-trained on TinyImageNet(Le and Yang, 2015). For each dataset, we conducted several tests using a different number of quantization bits (both for forward and backward passes) with the same training procedure. In addition to the work of Vorabbi et al.(Vorabbi et al., 2023a), in our experiments we kept different bitwidths for binary and non-binary layers because, as reported in Table 1, memory of trainable binary weights is predominant.

Hereafter we report some details about the dataset benchmarked and related CL protocols:

CORE50 (Lomonaco and Maltoni, 2017). It is a dataset specifically designed for Continuous Object Recognition containing a collection of 50 domestic objects belonging to 10 categories. The dataset has been collected in 11 distinct sessions (8 indoor and 3 outdoor) characterized by different backgrounds and lighting. For the continuous learning scenarios (NI, NC) we use the same test set composed of sessions #3, #7 and #10; the accuracy on test set is measured after each learning experience. The remaining 8 sessions are split in batches and provided sequentially during training obtaining 9 experiences for NC scenario and 8 for NI. No augmentation procedure has been implemented since the dataset already contains enough variability in terms of rotations, flips and brightness variation. The input RGB image is standardized and rescaled to the size of $128 \times 128 \times 3$.

CIFAR10 and CIFAR100 (Krizhevsky et al., 2009). Due to the lower number of classes, the NC scenario for CIFAR10 contains 5 experiences (adding

¹<https://docs.larq.dev/zoo/api/sota/>

2 classes for each experience) while 10 are used for CIFAR100. For both datasets the NI scenario is composed by 10 experiences. Similar to CORE50, the test set does not change over the experiences. The RGB images are scaled to the interval $[-1.0; +1.0]$ and the following data augmentation was used: zero padding of 4 pixels for each size, a random 32×32 crop and a random horizontal flip. No augmentation is used at test time.

On CORE50 dataset, we evaluated the three binary models reported below:

Quicknet and QuicknetLarge (Bannink et al., 2021). This network follows the previous works (Liu et al., 2018; Bethge et al., 2019; Martinez et al., 2020) proposing a sequence of blocks, each one with a different number of binary 3×3 convolutions and residual connections over each layer. Transition blocks between each residual section halve the spatial resolution and increase the filter count. QuicknetLarge employs more blocks and feature maps to increase accuracy. For Quicknet, latent replay memory has been set to quant.conv2d_16 layer by storing 1-bit activations; for QuicknetLarge the latent replay level is quant.conv2d_30. At this level (both for Quicknet and QuicknetLarge) activation has a dimensionality of $(7, 7, 256)$ and storing in the replay memory 1-bit activations leads to a considerable memory saving.

In contrast to the findings presented in (Vorabbi et al., 2023a), our study did not include the *Realtobinary* (Martinez et al., 2020) model, as it achieved notably lower accuracy levels that were not aligned with our research objectives and goals.

For CIFAR10 and CIFAR100 datasets, whose input resolution is 32×32 , we evaluated the following networks (pre-trained on Tiny Imagenet):

BiRealNet (Liu et al., 2018). It is a modified version of classical ResNet that proposes to preserve the real activations before the sign function to increase the representational capability of the 1-bit CNN, through a simple shortcut. Bi-RealNet adopts a tight approximation to the derivative of the non-differentiable sign function with respect to activation and a magnitude-aware gradient to update weight parameters. We used the instance of the network that uses *18-layers*². The latent replay layer has been set to add_12. At this level activation has a dimensionality of $(4, 4, 512)$.

²Refer to the following <https://github.com/liuzechun/Bi-Real-net> repository for all the details.

ReactNet (Liu et al., 2020). To further compress compact networks, this model constructs a baseline based on MobileNetV1 (Howard et al., 2017) and adds a shortcut to bypass every 1-bit convolutional layer that has the same number of input and output channels. The 3×3 depth-wise and the 1×1 point-wise convolutional blocks of MobileNet are replaced by the 3×3 and 1×1 vanilla convolutions in parallel with shortcuts in ReactNet³. As for Bi-Real Net, we tested the version of React Net that uses *18-layers*. The latent replay layer has been set to add_12 layer. At this level activation has a dimensionality of $(4, 4, 512)$.

In our experimental setup, we discovered that reducing the number of epochs in each learning experience had minimal impact on model accuracy. Consequently, we empirically set the number of epochs to 5, thus constraining the training time on-device platform. Across all classification tasks, we utilized the Cross Entropy loss function in conjunction with Stochastic Gradient Descent (SGD) as the optimizer. The former was chosen due to its simplicity in derivative computation when combined with the Softmax activation function. The latter was preferred for its computational efficiency, offering lower overhead compared to more complex algorithms like Adam (Kingma and Ba, 2014). In our experiments, the ratio of B_N to the batch size of the latent activations sampled from the replay memory is set at $1/4$. Both weight and activation binarization were performed during training, including both the first training experience and on-device stages. This choice requires the implementation of a quantized backward pass technique for all the non-differentiable functions, specifically the binarization functions (using Eq. 4 and 1). To assess model accuracy during on-device training, we developed the quantized backward steps for all layers employed by the previously described models.

Our experiments primarily concentrated on the NC scenario. As highlighted in (Pellegrini et al., 2020), the adoption of a latent replay memory did not significantly enhance model accuracy in the NI context. Moreover, the NC scenario more closely resembles real-world applications where the model’s recognition capability must be expanded to accommodate new, previously unknown classes.

4.1 Accuracy Comparison

To assess the accuracy of our proposed solution, we initiated our evaluation by comparing it with

³Refer to the following <https://github.com/liuzechun/ReactNet> repository for all the details.

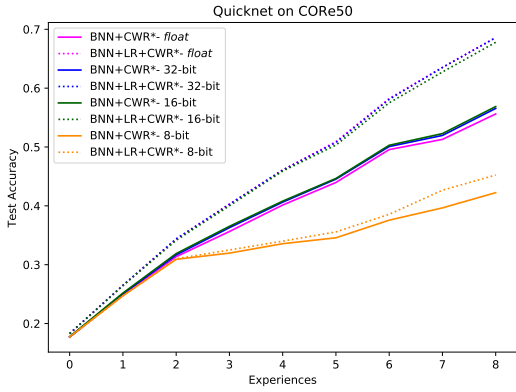


Figure 3: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CORE50 using *quick* model.

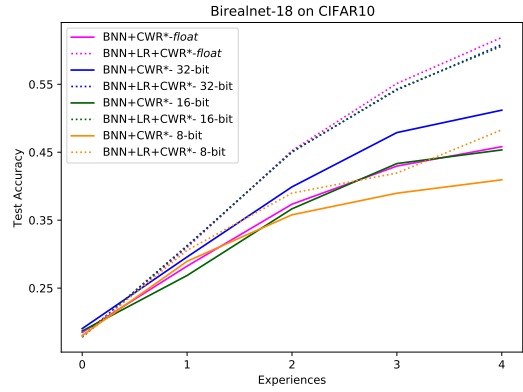


Figure 6: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CIFAR10 using *Birealnet* model.

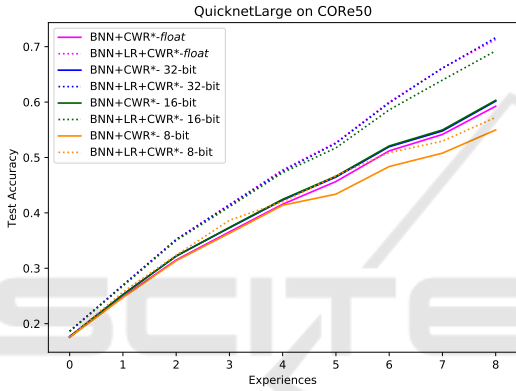


Figure 4: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CORE50 using *QuickNetLarge* model.

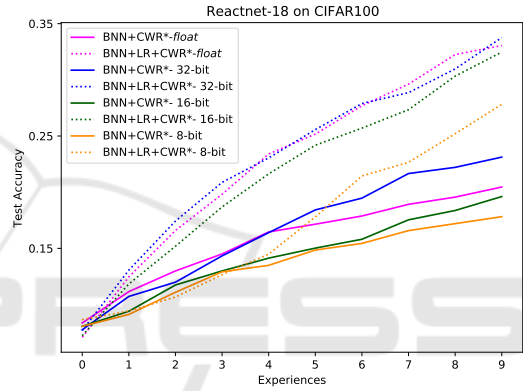


Figure 7: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CIFAR100 using *Reactnet* model.

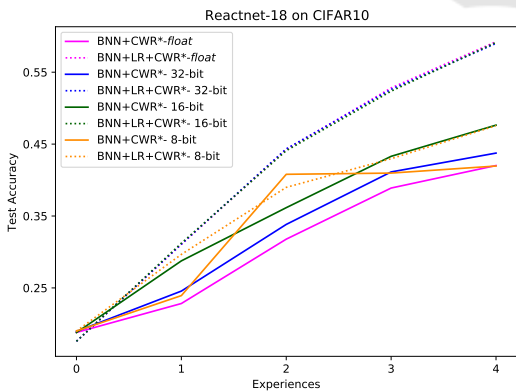


Figure 5: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CIFAR10 using *Reactnet* model.

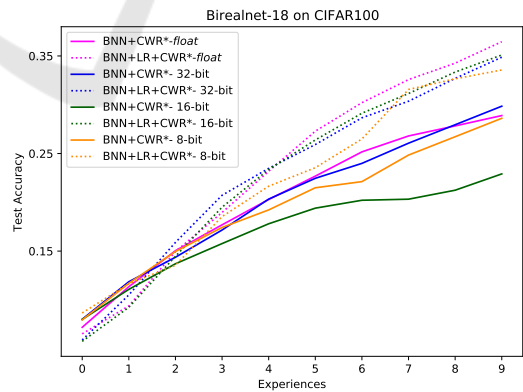


Figure 8: Accuracy comparison of our solution (BNN+LR+CWR*) with previous work BNN+CWR* (Vorabbi et al., 2023a) on CIFAR100 using *Birealnet* model.

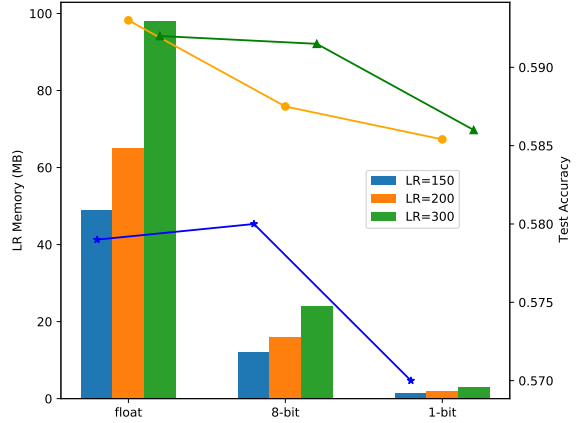
prior work, specifically BNN+CWR* (Vorabbi et al., 2023a), where only the final classification layer is trained on-device, without employing a replay memory. We conducted a series of tests with varying

quantization bitwidths for both forward and backward passes. In Fig. 3, 4, 5, 6, 7 and 8 we present accuracy comparisons between BNN+CWR* with the current method, denoted as **BNN+LR+CWR***,

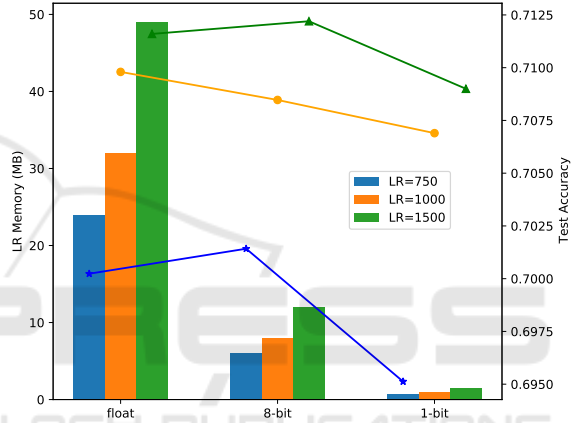
across different datasets: CORE50, CIFAR10 and CIFAR100. Each figure illustrates the performance improvement of the new method for all quantization settings tested, encompassing floating-point arithmetic, 32-bit, 16-bit and 8-bit quantized representations. It is noteworthy that, in this assessment, we applied the same quantization bitwidths (q_b) for both binary (q_b^{bin}) and non-binary ($q_b^{non-bin}$) layers during the backward pass, as BNN+CWR* does not distinguish these cases. The results consistently demonstrate that our BNN+LR+CWR* approach outperforms previous results, not only when using floating-point arithmetic but also for quantized implementations. This underscores the superior performance achieved by BNN+LR+CWR*. In our solution, we observed that employing $q_b = 8$ in BNN+LR+CWR* leads to a notable drop in accuracy compared to higher quantization bitwidth settings, aligning with the outcomes obtained by BNN+CWR*. This reaffirms the importance of using higher bitwidth representations during the backward pass to preserve model accuracy. For the experiments, we utilized $LR_{size} = 1500$ for CORE50, $LR_{size} = 300$ for CIFAR10 and $LR_{size} = 3000$ for CIFAR100 as our replay memory sizes.

4.2 Reducing Storage in Latent Replay

The storage requirements of the latent replay memory are closely interlinked with the bitwidths utilized to represent latent activations. As the bitwidths increase, so does the memory footprint of LR. In our approach we capitalize on the 1-bit activations inherent to BNNs to significantly mitigate the need for high-memory storage while maintaining a minimal accuracy gap, as depicted in Fig. 9. Our experiments demonstrate that BNN models can attain a minimal accuracy gap on both CIFAR10 and CORE50 datasets, even when adopting 1-bit latent activations for LR. This translates to a huge memory reduction of $32\times$ when compared to using floating-point latent activations. In our analysis, we considered various sizes for the LR memory, with 15, 20 and 30 elements allocated for each class. Importantly, we observed that the number of past samples in LR had a relatively minor impact on model accuracy, with the accuracy loss being within 1%. Utilizing 1-bit latent activations for LR opens the possibility to scale up applications to accommodate thousands of classes, as illustrated in Fig. 9, thanks to the substantial reduction in memory constraints achieved.



(a) Reactnet-18 on CIFAR10.

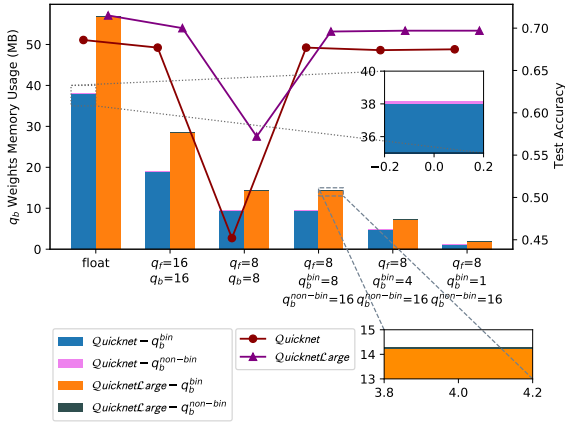


(b) Quicknet on CORE50.

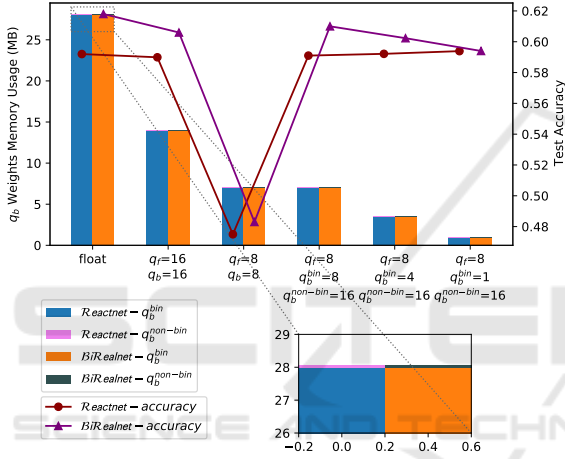
Figure 9: LR memory requirement using different quantization levels and corresponding test set accuracy on CIFAR10 (a) and CORE50 (b). We considered 15, 20 and 30 elements for each class inside LR; for case (a) we adopted Reactnet-18 model while in (b) we used Quicknet.

4.3 Splitting q_b in q_b^{bin} and $q_b^{non-bin}$

As highlighted in Table 1, the memory footprint of BNN weights is predominantly occupied by trainable binary weights, encompassing nearly 100% of the memory. Conventionally, a binary layer is trained using latent floating-point weights (Helweggen et al., 2019). However, if we were to replicate this approach on the device, it would result in a substantial increase in memory storage requirements during backpropagation stage, as it would require setting $q_b^{bin} = 32 - bit$. The quantization methodology proposed in Section 3.3 offers a potential solution to mitigate this constraint by reducing q_b to 8 bits. However, as depicted in Figure 10a and 10b, such a reduction in bitwidths would lead to a noticeable accuracy drop in the model. To address this challenge, we evalu-



(a) Quicknet and QuicknetLarge on CORE50.



(b) Reactnet-18 on CIFAR10.

 Figure 10: q_b memory requirement using different quantization bitwidths for backward layer on CORE50 (a) and CIFAR10 (b).

ated the impact of distinct quantization levels for binary weights (q_b^{bin}) and non-binary weights ($q_b^{non-bin}$). Specifically, we experimented with representing q_b^{bin} using both 4 bits and 1 bit. Our findings, as shown in Figure 10, indicate that 4-bit representation for binary layers does not introduce a substantial accuracy loss. Moreover, employing a 1-bit representation of weights during the back-propagation stage is feasible, as binary weights remain frozen during on-device learning. In this scenario, the model still effectively preserves accuracy. This latest result carries significant implications for on-device learning, as it simplifies the computational burden by requiring backward steps only for non-binary layers, primarily those employing $q_b^{non-bin} = 16$ - bits, as observed in our experiments.

4.4 Efficiency Evaluation

To demonstrate the applicability of our approach on real-world embedded boards, we provide an estimation analysis of the on-device performance. For this evaluation, we select two popular boards commonly used in the IoT paradigm, both based on the single-thread ARMv8 platform: Raspberry Pi 3B and Raspberry Pi 4B. Based on the efficiency analysis reported in (Bannink et al., 2021; Pellegrini et al., 2020), we report in Table 2 the inference and backward timings of our BNN+LR+CWR* method compared to a non-binary solution (using a Mobilenetv2) (Pellegrini et al., 2020): the results obtained adopting Mobilenetv2 rely on floating-point precision for layers from LR up to the classification head. The frozen backbone is quantized using 8-bit (latent activations are stored with 8-bit precision) and executed with Tensorflow-Lite. Instead, BNN+LR+CWR* employs Quicknet model with the following quantization setting: $q_f = 8, q_b^{bin} = 8, q_b^{non-bin} = 16$; the framework used to execute binary inference is LCE (Bannink et al., 2021). The image input size considered is 224×224 and the batch size is 1. Our empirical evaluation for backward pass shows that our BNN+LR+CWR* can achieve a minimum speedup of $2 \times$ compared to a non-binary solution. In our evaluation we consider the worst-case scenario for backward step by setting $q_b^{bin} = 8$; instead, by setting $q_b^{bin} = 1$, the speedup reported in the fifth column of Table 2 should improve significantly.

5 CONCLUSION

On-device training holds great potential in the realm of the IoT, as it can facilitate the widespread adoption of deep learning solutions. In this study, our primary focus was the implementation of Binary Neural Networks (BNNs) in combination with Continual Learning algorithms, an approach not yet fully investigated in the literature. In particular, we propose the use of the CWR* method with the support of a replay memory, implementing several customized quantization schemes tailored to alleviate memory constraints and computational bottlenecks during the back-propagation stage. Summarizing, experimental achievements in this work include the following:

- **Reduced Memory Usage:** we significantly reduced the memory storage required for replay memory by employing 1-bit latent activations, as opposed to the state-of-the-art approach that employs 8-bit precision. A limited storage requirement is a key element in addressing on-device

Table 2: Efficiency comparison of our method implemented on two different embedded boards, *i.e.* Raspberry Pi 3B and 4B, using Mobilenetv2 and Quicknet model. As shown, our solution achieves up to $2.2\times$ speedup on the same platform.

Model	Raspberry		Binary	Quantization		Forward	Backward	Speedup
	3B	4B		q_f	q_b			
Mobilenetv2 (Howard et al., 2017)	✓			8-bit	float	340	134	1.0×
Quicknet (Bannink et al., 2021)	✓		✓	1-bit	16-bit	160	55	2.0×
Mobilenetv2 (Howard et al., 2017)		✓		8-bit	float	225	90	1.0×
Quicknet (Bannink et al., 2021)		✓	✓	1-bit	16-bit	105	38	2.2×

training, especially with embedded systems with a limited storage capability.

- **Improved Model Accuracy:** we improve the accuracy obtained across different binarized backbones and the BNN+CWR* approach. Specifically, we reduce the gap in performance that commonly affects BNNs by introducing a latent replay approach as a safeguard against catastrophic forgetting.
- **Efficiency in Backpropagation:** we minimize the computational effort related to the backpropagation of the latent replay through a proper quantization scheme. In this manner, we combine the good performance of the model with limited computation requirements for the learning phase. This achievement, in combination with reduced memory usage, paves the way for future on-device and real-world training of learning systems.

A variety of future work can be planned based on the technological advancements introduced in this paper. For instance, we plan to effectively implement and optimize the approach proposed in this paper for the specific ARM CPUs, a popular family of processors often used in IoT devices. In addition, we envisage the possibility of exploiting their instruction set, including NEON extensions, to further optimize the proposed method in terms of computational load and efficiency.

REFERENCES

- Alshehri, F. and Muhammad, G. (2020). A comprehensive survey of the internet of things (iot) and ai-based smart healthcare. *IEEE Access*, 9:3660–3678.
- Banbury, C. R., Reddi, V. J., Lam, M., Fu, W., Fazel, A., Holleman, J., Huang, X., Hurtado, R., Kanter, D., Lokhtov, A., et al. (2020). Benchmarking tinyml systems: Challenges and direction. *arXiv preprint arXiv:2003.04821*.
- Banner, R., Hubara, I., Hoffer, E., and Soudry, D. (2018). Scalable methods for 8-bit training of neural networks. *Advances in neural information processing systems*, 31.
- Bannink, T., Hillier, A., Geiger, L., de Bruin, T., Overweel, L., Neeven, J., and Helwegen, K. (2021). Larq compute engine: Design, benchmark and deploy state-of-the-art binarized neural networks. *Proceedings of Machine Learning and Systems*, 3:680–695.
- Bethge, J., Yang, H., Bornstein, M., and Meinel, C. (2019). Back to simplicity: How to train accurate bnns from scratch? *arXiv preprint arXiv:1906.08637*.
- Cai, H., Gan, C., Zhu, L., and Han, S. (2020). Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33:11285–11297.
- Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*.
- Das, D., Mellempudi, N., Mudigere, D., Kalamkar, D., Avancha, S., Banerjee, K., Sridharan, S., Vaidyanathan, K., Kaul, B., Georganas, E., et al. (2018). Mixed precision training of convolutional neural networks using integer operations. *arXiv preprint arXiv:1802.00930*.
- Graffieti, G., Borghi, G., and Maltoni, D. (2022). Continual learning in real-life applications. *IEEE Robotics and Automation Letters*, 7(3):6195–6202.
- Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. (2015). Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR.
- Helwegen, K., Widdicombe, J., Geiger, L., Liu, Z., Cheng, K.-T., and Nusselder, R. (2019). Latent weights do not exist: Rethinking binarized neural network optimization. *Advances in neural information processing systems*, 32.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.
- Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of*

- the *IEEE conference on computer vision and pattern recognition*, pages 2704–2713.
- Jacob, B., Warden, P., and Guney, M. (2017). gemmlowp: a small self-contained low-precision gemm library. (2017). URL <https://github.com/google/gemmlowp>.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Le, Y. and Yang, X. (2015). Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3.
- Lin, J., Zhu, L., Chen, W.-M., Wang, W.-C., Gan, C., and Han, S. (2022). On-device training under 256kb memory. *Advances in Neural Information Processing Systems*, 35:22941–22954.
- Liu, Z., Shen, Z., Savvides, M., and Cheng, K.-T. (2020). Reactnet: Towards precise binary neural network with generalized activation functions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 143–159. Springer.
- Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., and Cheng, K.-T. (2018). Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proceedings of the European conference on computer vision (ECCV)*, pages 722–737.
- Lomonaco, V. and Maltoni, D. (2017). Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, pages 17–26. PMLR.
- Lomonaco, V., Maltoni, D., and Pellegrini, L. (2020). Rehearsal-free continual learning over small non-iid batches. In *CVPR Workshops*, volume 1, page 3.
- Martinez, B., Yang, J., Bulat, A., and Tzimiropoulos, G. (2020). Training binary neural networks with real-to-binary convolutions. *arXiv preprint arXiv:2003.11535*.
- Masana, M., Liu, X., Twardowski, B., Menta, M., Baganov, A. D., and Van De Weijer, J. (2022). Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533.
- Mohamed, E. (2020). The relation of artificial intelligence with internet of things: A survey. *Journal of Cybersecurity and Information Management*, 1(1):30–24.
- Nadalini, D., Rusci, M., Benini, L., and Conti, F. (2023). Reduced precision floating-point optimization for deep neural network on-device learning on microcontrollers. *arXiv preprint arXiv:2305.19167*.
- Nadalini, D., Rusci, M., Tagliavini, G., Ravaglia, L., Benini, L., and Conti, F. (2022). Pulp-trainlib: Enabling on-device training for risc-v multi-core mcus through performance-driven autotuning. In *International Conference on Embedded Computer Systems*, pages 200–216. Springer.
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, 113.
- Pellegrini, L., Graffieti, G., Lomonaco, V., and Maltoni, D. (2020). Latent replay for real-time continual learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10203–10209. IEEE.
- Qin, H., Gong, R., Liu, X., Bai, X., Song, J., and Sebe, N. (2020). Binary neural networks: A survey. *Pattern Recognition*, 105:107281.
- Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer.
- Ravaglia, L., Rusci, M., Nadalini, D., Capotondi, A., Conti, F., and Benini, L. (2021). A tinyml platform for on-device continual learning with quantized latent replays. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11(4):789–802.
- Ren, H., Anicic, D., and Runkler, T. A. (2021). Tinyol: Tinyml with online-learning on microcontrollers. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57.
- Vorabbi, L., Maltoni, D., and Santi, S. (2023a). On-device learning with binary neural networks. *arXiv preprint arXiv:2308.15308*.
- Vorabbi, L., Maltoni, D., and Santi, S. (2023b). Optimizing data-flow in binary neural networks.