2023

Energy-Aware AI-Driven Framework for Edge-Computing-Based IoT Applications

Muhammad Zawish
South East Technological University, Ireland

Nouman Ashraf
Technological University Dublin, Ireland, nouman.ashraf@tudublin.ie

Rafay Iqbal Ansari
Northumbria University, UK

See next page for additional authors

Energy-Aware AI-Driven Framework for Edge-Computing-Based IoT Applications

Muhammad Zawish, *Graduate Student Member, IEEE*, Nouman Ashraf, *Member, IEEE*, Rafay Iqbal Ansari, *Senior Member, IEEE*, and Steven Davy, *Member, IEEE*

***Abstract*—The significant growth in the number of Internet of Things (IoT) devices has given impetus to the idea of edge computing for several applications. In addition, energy harvestable or wireless-powered wearable devices are envisioned to empower the edge intelligence in IoT applications. However, the intermittent energy supply and network connectivity of such devices in scenarios including remote areas and hard-to-reach regions such as in-body applications can limit the performance of edge computing-based IoT applications. Hence, deploying state-of-the-art convolutional neural networks (CNNs) on such energy-constrained devices is not feasible due to their computational cost. Existing model compression methods, such as network pruning and quantization can reduce complexity, but these methods only work for fixed computational or energy requirements, which is not the case for edge devices with an intermittent energy source. In this work, we propose a pruning scheme based on deep reinforcement learning (DRL), which can compress the CNN model adaptively according to the energy dictated by the energy management policy and accuracy requirements for IoT applications. The proposed energy policy uses predictions of energy to be harvested and dictates the amount of energy that can be used by the edge device for deep learning inference. We compare the performance of our proposed approach with existing state-of-the-art CNNs and data sets using different filter-ranking criteria and pruning ratios. We observe that by using DRL-driven pruning, the convolutional layers that consume relatively higher energy are pruned more as compared to their counterparts. Thereby, our approach outperforms existing approaches by reducing energy consumption and maintaining accuracy.**

***Index Terms*—Artificial intelligence (AI), edge computing, energy efficiency, Internet of Things (IoT).**

I. Introduction

FIFTH-GENERATION (5G) wireless networks aim at supporting new applications and have given impetus to the evolution of the Internet of Things (IoT). The advancement toward sixth-generation (6G) networks will be empowered by solutions where IoT devices will play a significant role [1]. The rapid growth in the number of connected devices for applications, such as smart homes, smart cities, and wearable devices for healthcare has led to the utilization of edge computing, especially for scenarios with strict requirements of latency, reliability and energy efficiency [2], [3]. Moreover, the artificial intelligence (AI)-based solutions will play a key role in improving the key performance indicators (KPIs) in future 6G networks [4], [5].

European Telecommunications Standard Institute (ETSI) introduced the architecture that utilizes edge computing [6], leading toward the development of edge computing-assisted applications. The idea was to delegate the computation to the edge devices instead of undertaking them at the base station (BS). As a result, the data is processed near the source hence ensuring the security and low-latency decisions by saving the energy required for wireless transfer of data to the cloud. The edge devices with computing capability are in close vicinity of the end users, which allows them to realize several applications such as smart agriculture [5]. Despite the fact that due to edge computing, wearable devices are getting more realizable and inexpensive, AI-based applications that normally need greater processing resources may overload them with more data transfers and, as a result, increased energy consumption. However, as most edge devices (due to their mobile nature) are powered by batteries, their energy resources and operational hours are limited. Similarly, if enough battery power is not available for task transmission, compute performance may suffer. Thus, the computing capability and transmissions from edge devices heavily depend on the energy available at the device. This issue can be solved by using a bigger battery or charging it more frequently. The tiny size of edge or IoT devices, on the other hand, makes it difficult to equip them with bigger batteries or to recharge their batteries on a regular basis. Energy-harvesting technologies have been recognized as possible ways of increasing battery life and achieving energy-autonomous systems in order to meet these difficulties [7].

In several circumstances, the energy harvesting process may be affected by unpredictable extraneous environments, such as fluctuating solar irradiance in the case of solar energy harvesting. Furthermore, energy cannot be harvested at all times (such as at night), and the pace at which energy may be generated is restricted. The edge-computing enabled IoT devices with low energy consumption may incur long delays since the energy is not completely utilized to transmit at high data rates or compute at full capability, and the devices may miss recharging

Manuscript received 20 April 2022; revised 22 September 2022; accepted 29 October 2022. Date of publication 3 November 2022; date of current version 7 March 2023. This work was supported by the Science Foundation Ireland and the Department of Agriculture, Food, and Marine on behalf of the Government of Ireland VistaMilk Research Centre under Grant 16/RC/3835. *(Corresponding author: Muhammad Zawish.)*

Muhammad Zawish and Steven Davy are with the Walton Institute, South East Technological University, Waterford, X91 HE36 Ireland (e-mail: muhammad.zawish@waltoninstitute.ie; steven.davy@waltoninstitute.ie).

Nouman Ashraf is with the School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, D07 EWV4 Ireland (e-mail: nouman.ashraf@seecs.edu.pk).

Rafay Iqbal Ansari is with the Department of Computer and Information Sciences, Northumbria University, NE1 8QH Newcastle upon Tyne, U.K. (e-mail: rafay.ansari@northumbria.ac.uk).

Digital Object Identifier 10.1109/JIOT.2022.3219202

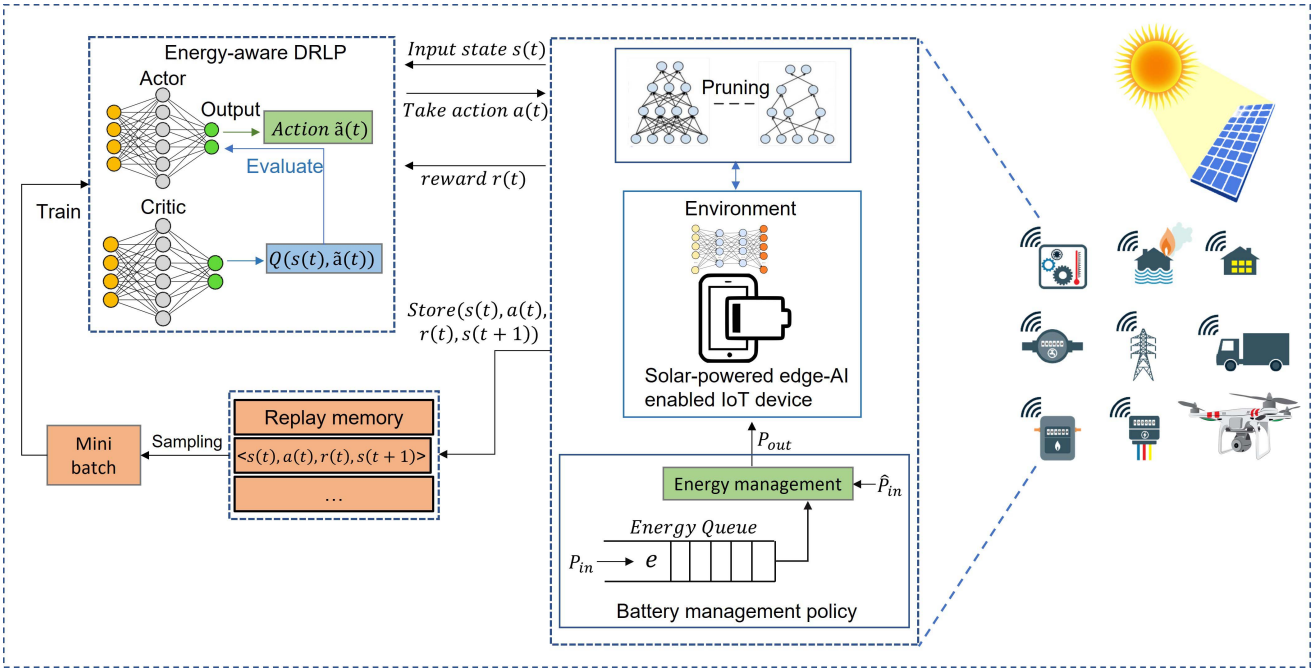

Fig. 1. Proposed energy-aware DRL-driven model compression and battery managemet scheme for Edge-AI enabled IoT applications.

opportunities owing to battery capacity restrictions. However, excessive energy usage may result in poor capability and lead to connections for brief periods of time [8]. Furthermore, total exhaustion of the battery restricts the computation of edge devices, which may be critical in particular applications. As a result, energy-neutral consumption strategies are necessary to avoid energy depletion, to ensure edge device performance for an extended period of time by balancing the aforementioned conflicting objectives, and maximizing the application performance.

To deal with the aforementioned scenario, this work focuses on energy-constrained IoT-enabled edge devices used for monitoring, especially in scenarios with intermittent energy supply and network connectivity. It is to be noted that the proposed approach applies to both energy-harvestable IoT devices and wireless-powered IoT devices. The utilization of edge devices such as wearables or unmanned aerial vehicles (UAVs) for monitoring has led to the development of several IoT applications [9]. Such edge devices can help to monitor the environment, capture images, and video, and process the data before sending it to the cloud [10]. Fig. 1 presents several edge-AI applications for IoT.

Convolutional neural networks (CNNs) are among the foremost AI algorithms for processing the raw data generated from the edge nodes and providing valuable insights. However, their computational complexity makes them practically infeasible for the aforementioned energy-constrained applications. Thus, state-of-the-art CNNs, such as VGG-16 [11] or ResNet [12] are required to be compressed according to intermittent requirements of edge devices. IoT-enabled edge nodes can be deployed either in isolated or outlying regions, such as in a rural environment, or are geographically spread in an area [3], [13], [14], thereby having limited network connectivity and energy supply. As a result, preventing increased energy usage and communication costs is a complex and demanding problem that needs to be addressed. In such scenarios, there is a need for an adaptive AI-driven scheme that can compress a CNN model dynamically based on the energy availability status. Deep reinforcement learning (DRL), a category of AI, has recently gained attention for utilization in systems having dynamic requirements. Thereby, this work focuses on a framework comprising DRL-driven adaptive model compression that can be helpful in realizing energy-constrained edge-empowered IoT applications.

Recent studies on CNN compression are mainly based on filter pruning [6], [15], [16], [17], [18], [19], [20], weight pruning [21], [22], [23], and quantization [22], [24]. These approaches either proposed a complex filter-ranking criteria or a joint multistage offline framework for reducing floating-point operations (FLOPs)/memory. These approaches are only suitable when a certain application has a fixed accuracy-complexity requirement. However, they do not cater for to the needs for applications of IoT, where resource requirement changes dynamically. In contrast, this work aims to propose a solution based on DRL, where any pretrained CNN can be pruned dynamically with the best accuracy and energy tradeoff. The proposed framework consists of two parts: 1) intelligent energy management and 2) energy-aware intelligent edge computing. The formal is based on modeling the battery of the edge device as a linear queue accommodating energy charges [25]. For this energy queue model, we use Model Predictive Control to design a control algorithm, which considers the prediction of energy to be harvested and regulates the output energy of the battery to a predetermined level [8]. This regulation of energy makes sure that the edge device always has energy available for its crucial operation. Our second algorithm takes the energy availability status parameter from the energy management system (EMS)

and accuracy requirement as input and prunes the CNN model iteratively until the desired accuracy-energy requirements are met. Then, the CNN model can be executed, for instance, on an energy-constrained wearable edge device deployed on or in the body of a human or an animal. The remainder of this article is organized as follows. We review the related literature in Section II, followed by the contribution and significance of this work. Sections III and IV describe the proposed approach and its performance evaluation, respectively, followed by a conclusion in Section V.

II. Related Work and Contributions

A. Compressing CNNs for Edge Devices

In recent years, various techniques have been proposed to optimize CNNs for resource-constrained edge devices. For example, network pruning [15], [16], [18], [22], [24], [26], quantization [22], and light-weight structure design [27] are a few techniques widely used for compressing or redesigning the CNNs. Network pruning has recently gained attention for reducing the complexity of CNNs without significantly degrading the performance. Several works propose techniques for identifying and removing unimportant weight connections to prune CNNs in an unstructured manner. For example, [21], [22], [23], demonstrated the significant compression on unstructured pruning of AlexNet, VGG-16, and ResNets. However, their proposed work underpins highly sparse models, which are complex to fine-tune and require additional time for hyperparameter optimization. In contrast, filter pruning was proposed to maintain the structured sparsity in CNNs. Therefore, it does not require any custom hardware or software for the execution, unlike weight pruning. For this reason, in this work, we focus on exploiting the idea of structured pruning using DRL.

B. CNN Pruning Schemes

Several techniques have been proposed to identify the unimportant filters in a convolutional layer, such as ℓ_1-norm [16], APoZ [15], Taylor expansion [17], and mean activation [26]. According to Hu et al. [15], eliminating those filters whose feature maps have the highest average percentage of zero activations leads to a compact model. Similarly, Li et al. [16] proposed to calculate the magnitude of filters using ℓ_1-norm, and removed filters with least ℓ_1-norm. However, these methods follow a one-shot pruning mechanism and yield a model with strict resource requirements. On the other hand, some studies proposed iterative pruning [6], [18], [19], where a model is compressed iteratively with a small pruning ratio to achieve the resource-accuracy tradeoff. Keeping in view the idea of hand-crafted complex filter ranking techniques, Mittal et al. [28] suggested that pruning filters randomly can also achieve similar performance as state-of-the-art ranking-based techniques. Thereby, a ranking framework does not essentially contribute to the performance of CNNs, and it is the plasticity of CNNs which helps them to recover after fine-tuning [28]. However, these works prune filters either from all layers or on an ad-hoc basis. As a result, this approach is not rationalized because each layer has a different level of complexity. In this work, we overcome it by compressing each layer with respect to its defined complexity, i.e., energy in our case.

C. DRL-Based CNN Compression

Recently, DRL-based techniques have emerged to automate the optimization of CNNs using policies, such as deep deterministic policy gradient (DDPG), Q learning, etc. [29], [30], [31]. Most of these techniques are focused on generating or searching the architecture of CNNs using the agent's search space. However, there is limited work available on DRL-based pruning (DRLP) mechanisms. For instance, Zhan et al. [24] recently proposed a two-stage framework based on joint optimization of CNNs using pruning and quantization. The authors use a DRL agent to sample the sparsity ratio for a certain layer based on an observation vector consisting of layer properties. However, the above works are only theoretically acceptable because they do not consider the dynamic resource requirements of mobile edge devices. In this work, we address the aforementioned problems by proposing a DRL-based automatic pruning scheme for CNNs where the model will be compressed according to the dynamic energy requirements of the device. Moreover, unlike previous studies, our DRL agent selects a particular layer for pruning based on its energy proportion with respect to the CNN's energy. The proposed technique can be particularly helpful in edge-empowered IoT applications, such as healthcare and agriculture, where intermittent energy and network conditions necessitate the need for a dynamic CNN compression approach.

D. Novelty and Significance of the Proposed Solutions

In this work, we propose a DRL-based CNN model compression to enable edge-AI in IoT applications. To summarize, following are the key contributions of this work.

1) We present an energy management strategy that uses the prediction of energy to be harvested and regulates the output power of the battery by using the Model Predictive Control method.
2) We propose a novel DRLP scheme, which prunes the model iteratively to meet the dynamic energy requirements of the energy-constrained edge devices for IoT applications.
3) We demonstrate the significance of the DRLP in two aspects: a) a CNN model is compressed automatically for immediate execution on edge devices based on the energy-accuracy criteria and b) the pruning mechanism considers the underlying energy of each layer in the CNN model and prunes a certain layer based on its proportion to overall model complexity.
4) We evaluate the performance of our approach through extensive simulation experiments and compare it with state-of-the-art methods on various filter-ranking mechanisms and pruning ratios using standard CNN models and benchmark data sets. It can be observed from

the results that our approach outranks other state-of-the-art methods on model pruning in terms of the energy-accuracy tradeoff.

5) We integrate energy management policy with our proposed DRLP method and perform extensive simulations to demonstrate the combined performance of the method by using realistic data of energy.

It is to be noted that the contribution of this work not only lies in proposing the use of energy-aware DRL for CNN compression but also in proposing a mechanism for adaptive on-demand compression of CNNs, which has not been explored in literature. In the subsequent sections, we present the methodology of our proposed framework and provide details on the energy consumption of CNNs and the energy-aware DRLP scheme. Moreover, the energy management scheme in this work is based on our previous work in [8]; however, in this work, instead of considering the case of IoT devices, we integrate this energy management strategy with edge computing-based IoT devices.

III. Proposed Energy-Aware Scheme

The framework for the energy-aware DRL-driven model compression and battery management is shown in Fig. 1. In the proposed model, we consider energy harvesting enabled edge devices, where these edge devices harvest ambient energy from the surrounding environment. The proposed framework consists of two parts: 1) a battery management policy and 2) an energy-aware DRLP scheme. Our energy management policy is responsible for monitoring the intermittent energy status of a certain edge device that is empowered with battery and solar power. The EMS regulates the available energy required for the execution of a CNN model on the edge devices for a certain IoT application [8]. The proposed DRLP model then considers the status of the energy system with intermittent energy supply and provides a compressed CNN model in return for execution. In contrast to existing compression schemes, our DRLP scheme prunes a pretrained CNN in an energy-aware manner such that each layer of CNN is pruned according to its weightage in the overall energy complexity of CNN. The DRL model needs to be pretrained on the GPU or a cloud server before deployment on the edge device. Then, any CNN model can be compressed using the proposed DRLP on the device itself without further training/fine-tuning and made available immediately for inference on edge, unlike offloading to the cloud. It has been reported in several studies that offloading inference tasks to the cloud requires orders of magnitude more energy than local inference on energy-harvesting edge devices [13], [23]. Therefore, the proposed scheme avoids computational offloading to cloud, hence reducing the energy expenditure required for it. In subsequent sections, we first present our energy management scheme. Second, we analyze the energy consumption of CNNs, followed by a comprehensive description of the proposed DRLP scheme. We demonstrate the DRLP scheme with an understanding of state, action, and reward functions curated for this task. The summary of notations is presented in Table I.

TABLE I
Summary of Notations

Notation	Definition
$P_{in}(t)$	Rate of the energy entering the battery
$P_{out}(t)$	Rate of the energy leaving the battery
$\hat{P}_{in}(t)$	Predicted rate of the energy to be harvested
M	Number of samples in the prediction horizon
$e(t)$	Energy stored in the battery
M_{FLOPs}	Total FLOPs for CNN model M with K $conv$ layers
M_{Memory}	Total memory for CNN model M with K $conv$ layers
M_{Energy}	Total energy for CNN model M with K $conv$ layers
C_{in}	Number of input channels
C_{out}	Number of input channels
Ω^2	Size of filters in the output layer
S_{out}	Size of feature maps in the output layer
E_{FLOP}	Energy required for a single FLOP
E_{access}	Energy required for DRAM access
P_r	Pruning ratio

A. Energy Management of Energy Harvestable Edge Device

In this section, we present our energy management framework for edge-enabled IoT devices. We model a battery of edge-enabled IoT devices as a queue that stores energy packets. To represent the energy queue, we adopt the queuing dynamics model shown below. The model relies on fluid flow principles

$$e(t) = P_{\text{in}}(t) - P_{\text{out}}(t) \tag{1}$$

where $P_{\text{in}}(t)$ denotes the rate of energy entering the battery, $e(t)$ represents the energy stored in the battery queue, and P_{out} denotes the rate of energy exiting the battery. The energy P_{out} is used for edge-enabled IoT device operation and affects the computation capability of the edge device. Our goal is to regulate the state of the above-defined queue or battery to a target level by controlling P_{out}. The control of the battery to a predetermined reference value ensures the device's continuous lifespan. In other words, P_{out} is controlled by pruning the CNN model. To regulate the battery to a predefined reference value, we use the Model Predictive Control to dictate the value of P_{out} at every instant. We formulate the problem as the following optimization problem:

$$P = \max_{P_{\text{out}}(k)} \quad \min_{k=1\cdots M} \quad P_{\text{out}}(k) \tag{2}$$

subject to

$$e(k+1) = e(k) + P_{\text{in}}(k) - P_{\text{out}}(k) \quad \text{for} \quad k = 1\cdots M \tag{3}$$

$$P_{\text{in}}(k) = \hat{P}_{\text{in}}(k) \quad \text{for} \quad k = 1\cdots M \tag{4}$$

$$e(\min) < e(k) < e(\max) \quad \text{for} \quad k = 1\cdots M \tag{5}$$

$$P_{\text{out}}(k) > 0 \quad \text{for} \quad k = 1\cdots M \tag{6}$$

where M in the number of samples in prediction horizon, $P_{\text{in}}(k)$ and P_{out} (k) are the corresponding discrete time variables for $P_{\text{in}}(t)$ and $P_{\text{out}}(t)$. $e(\min)$ and $e(\max)$ are the capacity constraints of the battery. It is to be noted that (3) represents the discrete time version of the battery model presented in (1) and the constraint in (4) defines the prediction of the energy to be harvested. Constraints in (5) are the bounds on minimum and maximum energy storage level of the battery such that charge stored in the battery can never be negative and always bounded by a maximum capacity value. Similarly, constraint in (6) makes sure that energy leaving the battery is always positive and battery supplies energy to the device.

B. Energy Consumption of CNNs

A standard CNN is composed of multiple convolutional and fully connected layers piled up in a vanilla shape. Among all layers of a CNN, convolutional layers perform the bulk of the computations, thus requiring higher energy for execution on low-powered IoT devices. A convolution operation is a basic element in a standard CNN whose kernels are defined by a 4-D tensor: $W \in \mathbb{R}^{C_{\text{in}} \times X \times Y \times C}$, where X and Y represent kernel's spatial dimensions, while C_{in} and C_{out} are the number of input and output channels, respectively. If $I \in \mathbb{R}^{C \times U \times V}$ are the input feature maps with a spatial dimension of U and V, then the output feature map f with the spatial dimension (x, y) can be calculated using

$$T(f, x, y) = \sum_{c=1}^{C} \sum_{x'=1}^{X} \sum_{y'=1}^{Y} I\left(c, x - x', y - y'\right) \cdot W\left(c, x', y', f\right). \tag{7}$$

Thereby, for a CNN model M with K convolutional layers, the FLOPs can be calculated using

$$M_{\text{FLOPs}} = \sum_{k=1}^{K} \left[C_{\text{in}}^{k} \times \left(\Omega^{k}\right)^{2} \times C_{\text{out}}^{k} \times S_{\text{out}}^{k} \right]. \tag{8}$$

Similarly, the memory consumption for a CNN model M with K convolutional layers can be calculated using

$$M_{\text{Memory}} = \sum_{k=1}^{K} \left[C_{\text{in}}^{k} \times \left(\Omega^{k}\right)^{2} \times C_{\text{out}}^{k} \times 4 \right]. \tag{9}$$

However, the energy requirement of a CNN is not only reflected by memory and FLOPs. In fact, the total energy consumption of a CNN is a sum of energy consumption in accessing the data and energy consumption in executing arithmetic operations. The former is a product of energy required for DRAM access (referred as E_{access}) in a 45-nm CMOS, i.e., 640 pJ [23] and model memory, while the latter is a product of energy required for a single FLOP (referred as E_{FLOP}), i.e., 2.3 pJ [23] and model's total FLOPs. Thus, the energy consumption for a CNN model M with K convolutional layers can be calculated using

$$M_{\text{Energy}} = \sum_{k=1}^{K} \left[\left(M_{\text{Memory}} \times E_{\text{access}}\right) + \left(M_{\text{FLOPs}} \times E_{\text{FLOP}}\right) \right] \tag{10}$$

where C_{in} and C_{out} are the number of input and output channels, respectively, while Ω^2 and S_{out} represent the size of filters and size of feature maps in the output layer, respectively. Moreover, M_{FLOPs}, M_{Memory}, and M_{Energy} denote the model's total FLOPs, memory, and energy. The M_{FLOPs}^{k}, M_{Memory}^{k}, and M_{Energy}^{k} denote the kth layer's FLOPs, memory, and energy, respectively. In this regard, we describe our pruning scheme in the subsequent section concerning the M_{Energy}^{k} and M_{Energy}.

C. DRLP: Deep Reinforcement Learning-Based Pruning

In this section, we describe our pruning scheme based on the actor-critic DRL approach. Unlike previous approaches, where the focus has been to obtain the pruning rate for each layer using DRL, we propose pruning of a particular layer based on its energy consumption using DRL. Hence, this approach is termed energy-aware pruning because it automatically caters to the individual layer's complexity instead of manually selecting a layer as done in existing methods. We leverage the actor-critic DRL approach that learns the optimal policy through deep neural networks (DNNs) and solves the layer-pruning problem with continuous action spaces. The actor-critic approach utilizes two DNNs; one serves as an actor network, while the other serves as a critic network. The actor network learns to take actions based on the policy gradient method and updates the actions given the evaluation from the critic network.

1) Actor, Critic, and Replay Memory: In our case, the actor DNN is responsible for producing labels on its output layer, which are modeled in continuous action space. The output on each node is quantized to either 0 or 1 for the corresponding layer, and the total output nodes are equal to the K convolutional layers for a particular CNN. The actions are generated using a parametrized function $\pi(s; \vartheta)$, where ϑ denotes the weight matrix of DNN. As the objective is to maximize the potential discounted reward $J(\pi)$, thus the weight ϑ is updated according to the direction of the objective function, $\nabla_{\vartheta} J(\pi_{\vartheta})$ [32]

$$\nabla_{\vartheta} J = \frac{\partial J}{\partial \pi} \frac{\partial \pi}{\partial \vartheta} = \nabla_{\tilde{a}} Q\left(s, \tilde{a}; \theta\right) \nabla_{\vartheta} \pi(s) \tag{11}$$

where $Q(s, \tilde{a}; \theta)$ is produced by critic which denotes the Q value for each state and action pair, θ is the weight matrix of the critic's DNN. Following equation is used to update the parameter ϑ for actor DNN:

$$\vartheta \leftarrow \vartheta + \alpha_a \nabla_{\vartheta} J \tag{12}$$

where α_a denotes the learning rate and is one of the hyperparameters of the actor. Using (12), the local optimum for the objective function $J(\pi)$ can be achieved over a number of iterative parameter updates. Meanwhile, the critic DNN keeps evaluating the actor's actions by monitoring the state-action Q value function $Q(s, \tilde{a}; \theta)$ that is responsible for updating the actor DNN's weight θ as shown in (11) and (12). The main objective of critic DNN is to minimize the temporal difference error $\delta(t)$ [33]

$$\delta(t) = -c(t+1) + \gamma Q\left(s(t+1), \tilde{a}(t+1); \theta\right) - Q\left(s(t), \tilde{a}(t); \theta\right) \tag{13}$$

where γ is the discount factor which defines the significance of future rewards. Similar to above equations, θ for critic DNN is updated in the direction of its temporal difference objective

$$\theta \leftarrow \theta + \alpha_c \nabla_{\theta} \delta(t) \tag{14}$$

where α_c acts as a learning rate hyperparameter for critic DNN. Moreover, during the training of actor and critic DNNs, training samples are stored in a finite-sized first-in-first-out (FIFO)-based replay memory. At each time epoch t, training samples $(s(t), a(t), c(t), s(t+1))$ are stored in this cache once the actions are made by the actor. Simultaneously, a mini-batch is sampled from the replay memory at each time epoch

to update weights ϑ and θ are updated to train the DNNs of actor and critic.

2) State, Action, and Reward: The proposed agent interacts with the environments, takes the state as an input, and produces action as an output that maximizes the reward. In our case, the environment is a CNN model. Thus, at a given time instant t, the proposed actor-critic DRL model takes state space $s(t)$ from the environment. After observing the environment at a time unit t, the DRL model takes the following set of parameters as a part of the state:

$$s(t) = \left\{ \left[I_k, C_{\text{in}}^k, C_{\text{out}}^k, M_{\text{Energy}}^k, R_k \right], \left[I_{k+1}, C_{\text{in}}^{k+1}, C_{\text{out}}^{k+1}, M_{\text{Energy}}^{k+1}, R_{k+1} \right], \ldots, \left[I_K, C_{\text{in}}^K, C_{\text{out}}^K, M_{\text{Energy}}^K, R_K \right] \right\}_t \tag{15}$$

where I_k, C_{in}^k, and C_{out}^k denote the index, input channels, and output channels of layer k, respectively. M_{Energy}^k is the current energy status of layer k, while R_k is the energy proportion of layer k in the overall CNN with K layers. The core idea behind providing these features of individual layers is to assist the agent in differentiating one convolutional layer from the other. Moreover, M_{Energy}^k and R_k particularly enhance the agent's ability to select a layer based on its underlying energy contribution to the model complexity.

Given the state space, the agent provides an action based on the policy defined above. The agent is required to decide whether a particular layer (k) should be selected for pruning or not in each iteration (i.e., $a_k \in \{0, 1\}$). Since there are 2^K possibilities of selection strategies, thus the number of output nodes in actor DNN can be calculated as 2^K. Thereby, the complexity of actor DNN exponentially increases with the number of layers (K) in a particular CNN. To overcome this, we model the actions in a scalar continuous action space which can be represented as $\tilde{a}_k \in [0, 1]$. Then, the selection decision for a certain layer is quantized to 0 or 1.

Finally, we model our reward function, which is inspired from [29] to evaluate the performance of the actor-critic compression policy. In [29], the reward function is synthesized in a way that it offers an incentive for not only maintaining the accuracy but also reducing the FLOPs or model size. However, the alternative reward function reward $= -$Error aggressively compresses the model without incentivizing the complexity reduction. Moreover, it has been empirically observed that Accuracy is directly proportional to the M_{Energy}. Thus, in our case, we model the reward function to incentivise both factors: reward $= -\text{Error} \cdot \log(M_{\text{Energy}})$. This reward function is constrained to not only $-$Error, but also the M_{Energy}. Hence, the model is compressed, taking care of the dynamic accuracy and energy budget required for a certain IoT application.

IV. Experiments and Results

In this section, we first discuss the experimental setup and analyze the impact of the DRLP scheme on energy, FLOPs, memory, and filters in each convolutional layer for different pruning ratios. Next, we evaluate the performance of our proposed DRLP model in terms of accuracy and energy tradeoffs over several iterations and pruning ratios with three filter pruning mechanisms, i.e., ℓ_1-norm, Taylor expansion, and random. Finally, we compare our model with state-of-the-art pruning techniques on different data sets and models.

A. Experimental Setup

We train our actor and critic networks with a learning rate of 1×10^{-3} and batch size of 100 on approximately 5000 episodes. Both DNNs are initialized with three fully connected hidden layers having sigmoid as an activation function. The input layer has a dimension equal to the number dimension of input states, while the output layer has a dimension equal to the number of layers in a CNN under observation. We perform all experiments using Keras deep learning API with Tensorflow as the backend. We pretrained the above models on an Nvidia Tesla K20 GPU machine with 8 GB of memory having Ubuntu 18.04 OS with Tensorflow version 1.15, Keras version 2.2.4, and Python version 3.6.12. All models are trained with data augmentation and follow the setup described in [18]. In the following sections, we mention some of the prominent CNN models and data sets which are considered to evaluate the proposed DRLP scheme. However, the approach can be extended to more similar CNNs and data sets.

1) DNN Models: We performed experiments on AlexNet, VGG-16, ResNet-18, and ResNet-32 which are usually considered the standard CNNs for benchmarking model compression approaches. AlexNet and VGG-16 were originally proposed for the ImageNet data set [11], [34], but several studies have reported their promising performance on CIFAR-100 and CIFAR-10 data sets. The topology of AlexNet is composed of five convolutional layers, followed by three fully connected layers. In comparison, VGG-16 is a slightly wider and deeper model consisting of 13 convolutional and three fully connected layers stacked on top of each other. In both models, ReLu activation is used in all layers except the last layer, which involves softmax as an activation function and represents the number of classes in a specific data set, e.g., 100 for CIFAR-100. We also consider ResNet-18 and ResNet-32, which are among the prominent models of the ResNet family. In general, both ResNet-18 and ResNet-32 begin with a convolutional layer followed by several parallel branches of layers along with skip connections and a fully connected layer in the end. However, ResNet-18 consists of eight residual units, and ResNet-32 consists of 15 units, and each residual unit is made up of two convolutional layers.

2) Data Sets: We use two data sets from CIFAR for the evaluation, namely, CIFAR-10 and CIFAR-100, which are the standard data sets for benchmarking pruning techniques. Both data sets consist of 60 000 digital images of size 32×32, with 50 000 images as a training set and 10 000 images as a test set. In CIFAR-10, there are ten classes, and each class has 5000 training and 1000 test images, while CIFAR-100 has 100 distinct classes, and for each class, there are 500 train images and 100 test images. In the subsequent, we present the performance evaluation of our proposed DRLP scheme.

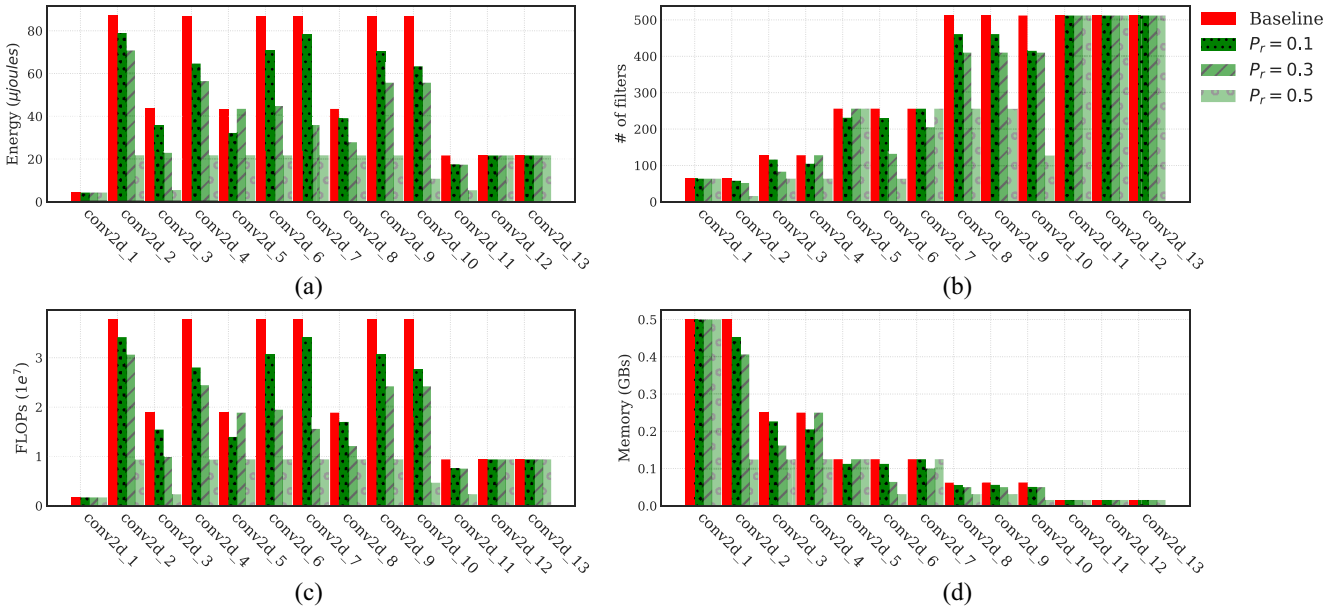

Fig. 2. (a) Energy consumption, (b) filters, (c) FLOPs, and (d) memory over 13 conv layers of VGG-16.

B. Performance Evaluation

The aim of our DRLP scheme is to automatically produce a compressed model given the dynamic energy requirements of edge devices in IoT. Hence, to validate the objective of our DRL model in terms of energy-aware pruning, we evaluate it on VGG-16 on CIFAR-10, as shown in Fig. 2. We perform pruning up to ten iterations with pruning ratios $P_r = \{0.1, 0.3, 0.5\}$. Fig. 2(a)–(d) show the layer-wise energy, filters, FLOPs, and memory reduction for VGG-16, respectively. The x-axis shows the indices of the convolutional layers in the corresponding CNN, while the y-axis shows: 1) the energy consumption; 2) number of filters; 3) FLOP count; and 4) memory consumption for each layer.

In contrast to typical pruning methods where layers are pruned according to the proportion of filters they carry, our proposed model prunes a certain layer based on the amount of energy it consumes. More importantly, a layer with more filters does not necessarily consume more energy. In fact, the energy consumption of each layer is constituted of not only the FLOP counts but also the memory access. Therefore, the energy required for accessing data from the DRAM for a certain operation is higher than the energy required for the operation itself. As a result, DRAM access for both feature maps and filter weights dominates the overall energy consumption of a CNN. It can be observed in Fig. 2 that a few layers consume the same amount of energy regardless of an unequal number of filters. This inconsistent distribution of energy and filters over layers is due to the impact of intermediate feature maps among layers. For instance, there is a slight reduction in the energy of conv_2d_11 despite the negligible reduction in filters. This type of scenario justifies the aforementioned arguments as energy is reduced only due to the reduction in FLOPs of the corresponding layer, which is a result of the change in feature maps from the preceding layer.

We also evaluate the DRL model with different filter importance measure techniques. The proposed model is able to work with a wide range of filter ranking techniques; however, we evaluate it on following three state-of-the-art procedures.

ℓ_1-norm [16]: In this method, the relative importance of a certain filter F in each layer is measured by summing its absolute weights $\sum |F|$ which results in its ℓ_1-norm ($\|F\|_1$). In the next step, filters are sorted according to ℓ_1-norm value, and a percentage of filters with the smallest value are pruned.

Taylor Expansion [17]: For a training data set D and loss function C, this method aims to prune a model in such a way that the change in loss $|\Delta C|$ is minimal once the model has maximum nonzero feature maps. The $|\Delta C|$ after eliminating $h_{i,j}$ feature maps (i.e., obtained from parameters i, j) can be calculated using

$$|\Delta C(h_{i,j})| = |C(D, h_{i,j} = 0) - C(D, h_{i,j})| \tag{16}$$

where $C(D, h_{i,j})$ and $C(D, h_{i,j} = 0)$ are losses with and without considering feature maps $h_{i,j}$, respectively. The loss with $h_{i,j}$ included can be calculated with

$$C(D, h_{i,j} = 0) = C(D, h_{i,j}) - \frac{\delta C}{\delta h_{i,j}} h_{i,j} + R_1(h_{i,j} = 0)$$

where $R_1(h_{i,j} = 0)$ is the first-degree remainder of the Taylor expansion. We can ignore R_1, since this method considers only the first degree estimation, thus, (16) can be transformed as follows:

$$|\Delta C(h_{i,j})| = \left| C(D, h_{i,j}) - \frac{\delta C}{\delta h_{i,j}} h_{i,j} - C(D, h_{i,j}) \right| = \left| \frac{\delta C}{\delta h_{i,j}} h_{i,j} \right|$$

where $|\Delta C(h_{i,j})|$ denotes the importance of filter $w_{i,j}$ associated with feature map $h_{i,j}$, and both $(\delta C/\delta h_{i,j})$ and δC can

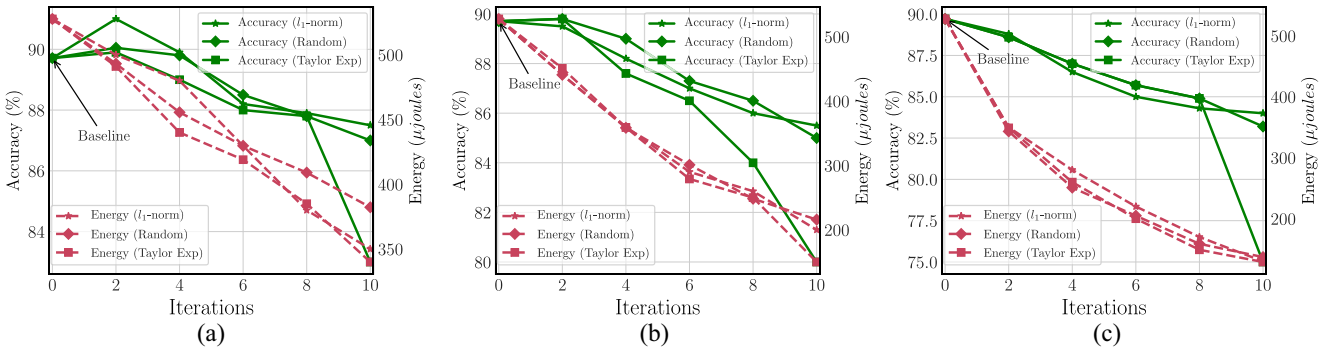

Fig. 3. Energy-accuracy tradeoff on ten pruning iterations and (a) $P_r = 0.1$, (b) $P_r = 0.3$, and (c) $P_r = 0.5$ for AlexNet on CIFAR-100.

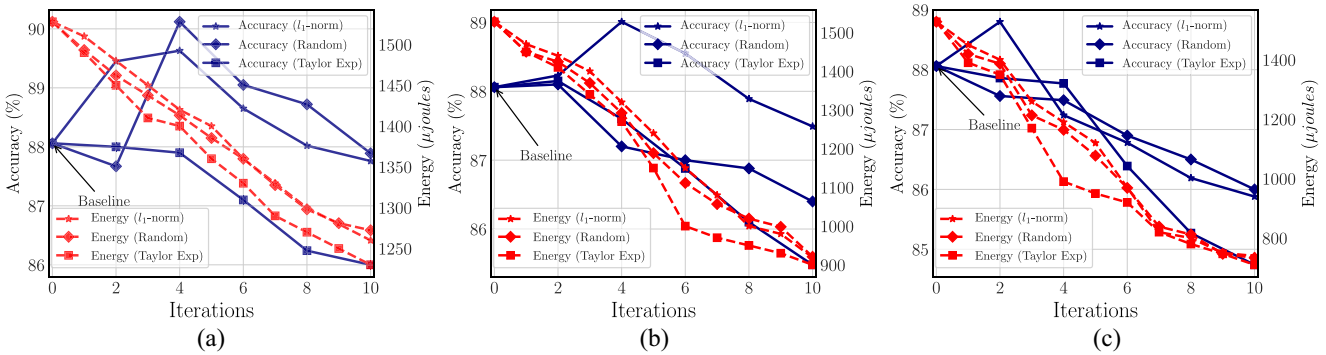

Fig. 4. Energy-accuracy tradeoff on ten pruning iterations and (a) $P_r = 0.1$, (b) $P_r = 0.3$, and (c) $P_r = 0.5$ for VGG-16 on CIFAR-10.

be obtained during the feed-forward and back-propagation of CNN by providing sufficient samples.

Random Pruning [28]: In this approach, filters are neither ranked nor hold any kind of importance metric to get eliminated from the layers. However, as opposed to ranking-based techniques, this method encourages randomly pruning a required percentage of filters from any layer without additional calculations.

We evaluate the effectiveness of aforementioned approaches on our compression scheme by pruning the AlexNet and VGG-16 up to ten iterations with pruning ratios $P_r = \{0.1, 0.3, 0.5\}$ as shown in Figs. 3 and 4, respectively. For Fig. 3(a), after soft pruning with $P_r = 0.1$, there is 33.6%, 27.5%, and 35.5% energy reduction and 2.4%, 3%, and 7.4% accuracy loss using ℓ_1-norm, random, and Taylor expansion approaches, respectively. Similarly, in Fig. 3(c), with a relatively higher pruning ratio $P_r = 0.5$, we observe 75.3%, 74%, and 75.5% energy reduction and 6.3%, 7.2%, and 16.3% accuracy loss using ℓ_1-norm, random, and Taylor expansion approaches, respectively. This behavior is consistent for the VGG-16 as well, which can be seen in Fig. 4(a)–(c). For the proposed scenario, it is obvious that ℓ_1-norm and random schemes are efficient, and Taylor expansion failed to minimize the accuracy loss. However, ℓ_1-norm involves some calculations that essentially increase the compression scheme's overall computation time. In contrast, randomly pruning does not incur additional computations but still performs consistently with ℓ_1-norm. Therefore, it can be concluded that the proposed DRL model should incorporate either random pruning or ℓ_1-norm (if latency can be tolerated) to efficiently tradeoff between energy and accuracy.

C. Comparison With State-of-the-Art

We compare our approach with state-of-the-art pruning approaches on two types of networks: 1) simple CNNs (AlexNet on CIFAR-100 and VGG-16 on CIFAR-10) and 2) Residual networks (ResNet-18 on CIFAR-10 and ResNet-32 on CIFAR-100). We use the same environment, preprocessing, and hyper-parameter tuning to get the baseline accuracy and energy. To reproduce [6], [18], [19], we follow the given guidelines and prune the model without fine-tuning it to synchronize with the proposed scheme. Thus, the accuracy can be different from the one mentioned in the original papers, as they all fine-tune the networks once they are pruned, which is not the case in our scenario.

1) VGG-16 on CIFAR-10: As shown in Fig. 5(a), the baseline accuracy for VGG is 88.06%, and the original model consumes approximately 1520 μJ. Our approach achieves the best tradeoff between loss in accuracy and energy reduction with ℓ_1-norm and $P_r = 0.3$ by reducing 40% energy and losing only 0.6% accuracy. However, [6], [18] reduce relatively more energy but fail to gain any improvement in accuracy. Moreover, [19] loses 1.2% accuracy, which is relatively lesser than [6], [18] but can only reduce 11% energy.

2) AlexNet on CIFAR-100: The baseline accuracy for AlexNet is 89.71%, and the uncompressed model consumes approximately 527 μJ as shown in Fig. 6(a). Reference [6] is able to reduce 72% of model energy aggressively as compared

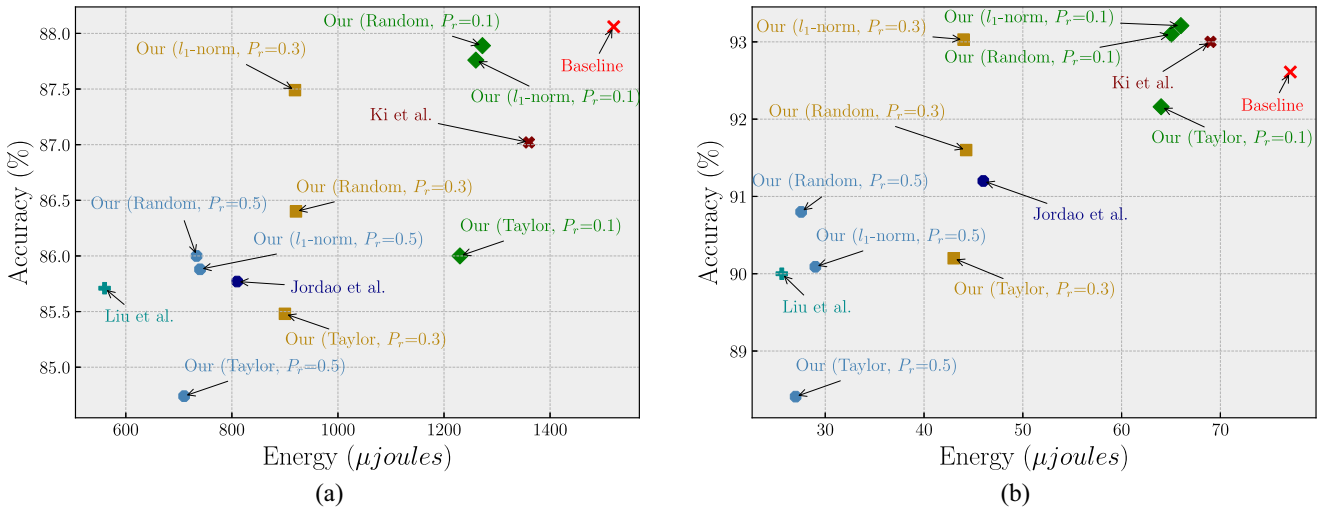

Fig. 5. Comparison of DRLP with state-of-the-art on (a) VGG-16-CIFAR-10 and (b) ResNet-18-CIFAR-10.

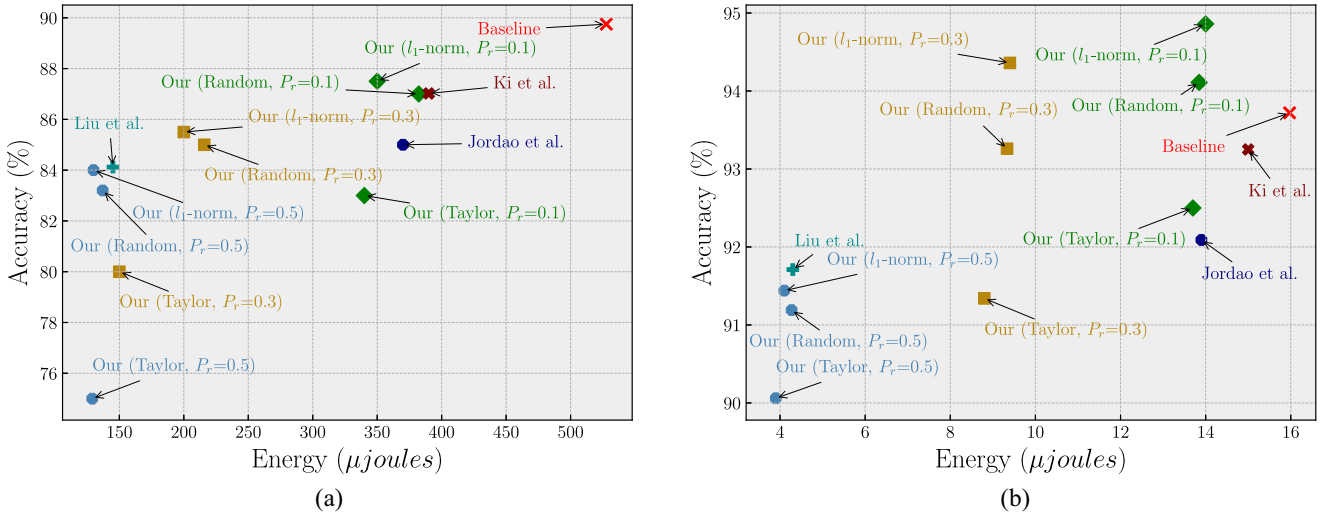

Fig. 6. Comparison of DRLP with state-of-the-art on (a) AlexNet-CIFAR-100 and (b) ResNet-32-CIFAR-100.

to [18] and [19], but in return, it loses 6.5% of accuracy. In contrast, our approach with ℓ_1-norm and $P_r = 0.5$ is able to reduce 75.3% of energy with only 6.3% of accuracy drop. Similarly, with random pruning and $P_r = 0.1$, our approach performs better than [19] by reducing 27.5% of energy with a minimal drop of 3% in accuracy.

3) ResNets on CIFAR-10 and CIFAR-100: We choose two well-known models from the ResNet family, i.e., ResNet-18 and ResNet-32, trained on CIFAR-10 and CIFAR-100, respectively. In general, pruning ResNets is challenging due to their architecture comprising several parallel branches and skip connections. Therefore, when dealing with such complex architectures, there are some requirements to adhere to as a result of interlayer dependencies.

For this reason, our DRL model does not select any layer with shortcut connections for pruning in order to maintain the structure in ResNets. The baseline ResNet-18 model has 92.61% accuracy and it consumes approximately 77 μJ as shown in Fig. 5(b). For ResNet-18, our approach achieves best tradeoff by gaining 0.45% of accuracy and reducing 43% of energy using ℓ_1-norm with $P_r = 0.3$. On the other hand, among state-of-the-art approaches, only [19] is able to gain 0.4% of accuracy but with minimal energy reduction. Moreover, the original ResNet-32 model has 93.72% accuracy and consumes approximately 16 μJ of energy, as shown in Fig. 6(b).

Similarly, for ResNet-32, the best tradeoff is achieved by our model with 0.7% accuracy gain and 40% energy loss using ℓ_1-norm with $P_r = 0.3$. In contrast, none of the state-of-the-art approaches could gain accuracy after reducing the energy. However, [6] is able to reduce the significant energy for both ResNet-18 and ResNet-32 but slightly lower than ours with random pruning and $P_r = 0.5$. It is clear that at a certain stage, our approach loses more accuracy by trading off the energy consumption, which is consistent with state-of-the-art studies. Therefore, as opposed to state-of-the-art, our approach can be used with different ranking criteria and pruning ratios to achieve the desired level of accuracy and complexity. For example, based on the above discussion, if a certain application can not tolerate a significant loss in accuracy but needs a reduction in energy, then either ℓ_1-norm or random criteria can be used with a small P_r, such as 0.1 or

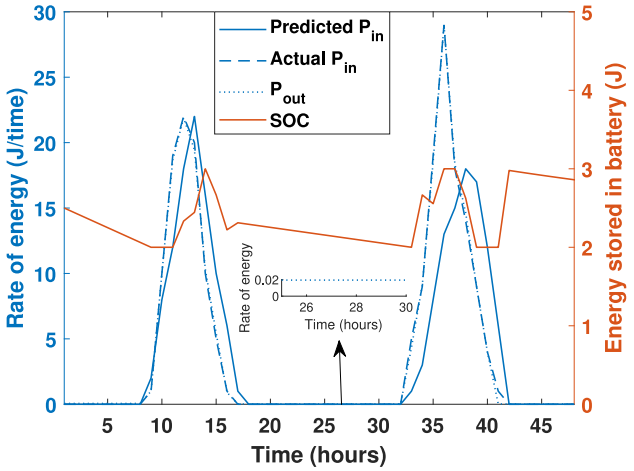

Fig. 7. Performance of energy management scheme.

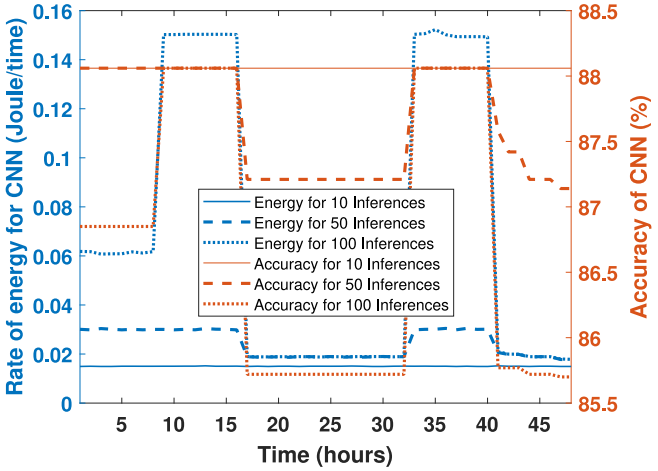

Fig. 8. Energy-accuracy tradeoff for VGG-16 over 10, 50, and 100 inferences based on P_{out}.

0.2. Alternatively, suppose the requirement of the application is to lower the energy footprint without taking care of accuracy. In that case, the aggressive pruning can be done using relatively larger P_r, such as 0.4 or 0.5 with suitable ranking criteria such as ℓ_1-norm/random in our case.

D. Performance Evaluation of Proposed DRLP Scheme in Presence of EMS

In this section, we evaluate the performance of the proposed DRLP in the presence of our energy management scheme. In particular, we demonstrate that our proposed energy-aware DRLP algorithm is efficient in adjusting the CNN model based on the time-varying energy constraints which arise due to the intermittent nature of harvestable energy. As discussed in the previous sections, for energy management, we model the battery of the IoT device as a linear buffer, which stores packets of the harvested energy. For the simulations, we are using realistic data for solar irradiance [7]. We use the Markov chain to predict the amount of energy to be harvested [7]. Fig. 7 depicts the energy predictions for 48 h, where 0 h represents midnight. Based on these energy predictions, our MPC-based energy management controller regulates the output energy of the battery (P_{out}), which is used for the operation of IoT devices, including the CNN model inferences. Fig. 7 shows the time evolution of P_{out}, which was dictated by our energy management scheme and the corresponding residual energy of the battery. It can be observed that based on the harvestable energy predictions; our scheme is able to regulate the P_{out} in such a way that residual energy (State of charge: SOC) in the battery is always above a predefined reference value even for the time instants when there is no harvestable energy available. This reference level was set to 2 J in our simulations. Moreover, P_{out} is always greater than zero, as shown in the magnifying window of Fig. 7. This P_{out} is the amount of energy that is utilized for CNN-based inferences. As shown in Fig. 7, the P_{out} is time varying in nature, which will affect the performance of the CNN model while performing the inference tasks. More precisely, for the particular time instance where P_{out} is below the energy required for the CNN model, it will fail to perform the given inference tasks. Therefore, to overcome this, our proposed energy-aware DRLP scheme is able to compress the model so it can adapt according to the time-varying nature of P_{out} at the cost of accuracy. To demonstrate the effectiveness of the proposed approach, we use the VGG-16 as the baseline CNN model, which is used to perform the inference tasks. The DRLP scheme prunes the VGG-16 model iteratively at every time instance until the energy required for a certain number of inferences is less than P_{out}. In Fig. 8, we show the tradeoff between energy and accuracy of VGG-16 over the 10, 50, and 100 inference tasks. It can be seen that the energy required for performing ten inference tasks by VGG-16 does not exceed the P_{out}; hence the accuracy remains unaffected. In contrast, the energy required for 50 and 100 inference tasks exceeds the P_{out} at time instances where the harvested energy is low. As a result, the number of inference tasks can be reduced at time instances where P_{out} is low and can be increased at time instances where P_{out} is high.

V. Conclusion

Edge computing is emerging as a prominent technique for several IoT applications. However, edge devices that are responsible for the execution of tasks are usually resource-constrained, hence cannot afford the deployment of raw CNN models. Therefore, this motivates a need for compressing the CNNs before deployment. The existing techniques to reduce CNNs complexity work on strict resource requirements and are not suitable in environments with intermittent energy supply. In this work, we considered an energy-aware AI-driven framework for edge-computing-based IoT applications. The proposed DRLP scheme can adapt according to the availability of energy. It can compress the model on the device dynamically to provide the best energy-accuracy tradeoff. Our results showed that the proposed DRLP outranks other state-of-the-art approaches on model pruning in terms of energy-accuracy tradeoff. In the future, we intend to consider the combined performance of our proposed scheme in the presence of an EMS.

References

[1] S. Liao, J. Wu, J. Li, and K. Konstantin, "Information-centric massive IoT-based ubiquitous connected VR/AR in 6G: A proposed caching consensus approach," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5172–5184, Apr. 2021.

[2] M. Wang, L. Zhu, L. T. Yang, M. Lin, X. Deng, and L. Yi, "Offloading-assisted energy-balanced IoT edge node relocation for confident information coverage," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4482–4490, Jun. 2019.

[3] J. Chen, W. Wang, Y. Zhou, S. H. Ahmed, and W. Wei, "Exploiting 5G and blockchain for medical applications of drones," *IEEE Netw.*, vol. 35, no. 1, pp. 30–36, Jan./Feb. 2021.

[4] H. Sami, H. Otrok, J. Bentahar, and A. Mourad, "AI-based resource provisioning of IoE services in 6G: A deep reinforcement learning approach," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 3, pp. 3527–3540, Sep. 2021.

[5] M. Zawish et al., "Towards on-device AI and blockchain for 6G enabled agricultural supply-chain management," 2022, *arXiv:2203.06465*.

[6] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2736–2744.

[7] N. Ashraf, M. Faizan, W. Asif, H. K. Qureshi, A. Iqbal, and M. Lestas, "Energy management in harvesting enabled sensing nodes: Prediction and control," *J. Netw. Comput. Appl.*, vol. 132, pp. 104–117, Apr. 2019.

[8] N. Ashraf, A. Hasan, H. K. Qureshi, and M. Lestas, "Combined data rate and energy management in harvesting enabled tactile IoT sensing devices," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3006–3015, May 2019.

[9] O. Salem, K. Alsubhi, A. Shaafi, M. Gheryani, A. Mehaoua, and R. Boutaba, "Man in the middle attack mitigation in Internet of Medical Things," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 2053–2062, Mar. 2022.

[10] X. Yuan, J. Chen, K. Zhang, Y. Wu, and T. Yang, "A stable AI-based binary and multiple class heart disease prediction model for IoMT," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 2032–2040, Mar. 2022.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[13] G. Gobieski, B. Lucia, and N. Beckmann, "Intelligence beyond the edge: Inference on intermittent embedded systems," in *Proc. 24th Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2019, pp. 199–213.

[14] F. A. Dharejo et al., "FuzzyAct: A fuzzy-based framework for temporal activity recognition in IoT applications using RNN and 3D-DWT," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 11, pp. 4578–4592, Nov. 2022.

[15] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," 2016, *arXiv:1607.03250*.

[16] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient ConvNets," 2016, *arXiv:1608.08710*.

[17] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–17.

[18] A. Jordao, F. Yamada, and W. R. Schwartz, "Deep network compression based on partial least squares," *Neurocomputing*, vol. 406, pp. 234–243, Sep. 2020.

[19] S.-K. Yeom et al., "Pruning by explaining: A novel criterion for deep neural network pruning," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107899.

[20] M. Zawish, S. Davy, and L. Abraham, "Complexity-driven CNN compression for resource-constrained edge AI," 2022, *arXiv:2208.12816*.

[21] N. Lee, T. Ajanthan, and P. H. Torr, "SNIP: Single-shot network pruning based on connection sensitivity," 2018, *arXiv:1810.02340*.

[22] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2015, *arXiv:1510.00149*.

[23] S. Han et al., "EIE: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 243–254, 2016.

[24] H. Zhan, W.-M. Lin, and Y. Cao, "Deep model compression via two-stage deep reinforcement learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2021, pp. 238–254.

[25] N. Ashraf, N. Christofides, and M. Lestas, "Combined data rate and energy management in wireless sensor networks with energy harvesting capability," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2018, pp. 6717–6722.

[26] A. Polyak and L. Wolf, "Channel-level acceleration of deep face representations," *IEEE Access*, vol. 3, pp. 2163–2175, 2015.

[27] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[28] D. Mittal, S. Bhardwaj, M. M. Khapra, and B. Ravindran, "Studying the plasticity in deep convolutional neural networks using random pruning," *Mach. Vis. Appl.*, vol. 30, no. 2, pp. 203–216, 2019.

[29] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "AMC: Automl for model compression and acceleration on mobile devices," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–800.

[30] Z. Zhong et al., "BlockQNN: Efficient block-wise neural network architecture generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2314–2328, Jul. 2021.

[31] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8697–8710.

[32] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.

[33] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Autom. Control*, vol. 42, no. 5, pp. 674–690, May 1997.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

Muhammad Zawish (Graduate Student Member, IEEE) received the B.E. degree in computer systems from Mehran UET, Jamshoro, Pakistan, in 2018. He is currently pursuing the Ph.D. degree in computer science with the Walton Institute, South East Technological University, Waterford, Ireland, under SFI VistaMilk Project.

Later, he worked as a Lab Engineer with the Department of Software Engineering, PAFKIET, Karachi, Pakistan. His research interests are deep learning for edge computing-based IoT applications.

Nouman Ashraf (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Department of EE, COMSATS University, Islamabad, Pakistan, in 2012 and 2015, respectively, and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, Frederick University, Nicosia, Cyprus, in 2019.

He has worked on the project VISORSURF (FETOPEN) with the University of Cyprus, Nicosia, and VistaMilk with Waterford Institute of Technology, Waterford, Ireland. He is currently working with Technological University Dublin, Dublin, Ireland, as an Assistant Lecturer. His research interests include the application of control theory for algorithm designing for emerging intelligent networks.

Rafay Iqbal Ansari (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Department of Computer Science and Engineering, Frederick University, Nicosia, Cyprus, in 2019.

He is currently working as a Senior Lecturer with the Department of Computer and Information Science, Northumbria University, Newcastle upon Tyne, U.K. His research interests lie in optimal resource allocation, 5G device-to-device networks, intelligent surfaces, millimeterWave networks, and public safety networks.

Steven Davy (Member, IEEE) received the B.A. degree (Mod Hons.) in computer science from Trinity College Dublin, Dublin, Ireland, in 2003, and the Ph.D. degree in computer science from Waterford Institute of Technology, Waterford, Ireland, in 2008.

He was the Head of Programmable Autonomous Systems, Walton Institute for Information and Communication Systems Science, Carriganore, Ireland. Also, he is a Funded Investigator on SFI VistaMilk Research Center and SFI CONNECT Research Center. He has lead more than 50 research projects funded under national and EU programmes. His research interests include distributed systems, network resource optimization and management, programmable autonomous systems and energy efficient artificial intelligence applied computer vision tasks.