Poster Abstract: DVFO: Dynamic Voltage, Frequency and Offloading for Efficient AI on Edge Devices

Ziyang Zhang
Harbin Institute of Technology
Harbin, China
zhangzy@stu.hit.edu.cn

Yang Zhao
Harbin Institute of Technology
Shenzhen, China
yang.zhao@hit.edu.cn

Jie Liu
Harbin Institute of Technology
Shenzhen, China
jieliu@hit.edu.cn

ABSTRACT

Due to resource constraints, it is challenging to optimize the inference performance in terms of energy consumption and latency on edge devices. In this paper, we leverage both the dynamic voltage frequency scaling (DVFS) technique and edge-cloud collaborative inference to minimize the overall energy consumption. We propose a deep reinforcement learning (DRL)-based method called DVFO to jointly optimize 1) CPU, GPU and memory frequencies, and 2) the ratio of offloaded feature maps in edge-cloud collaboration. Preliminary experimental results show that DVFO reduces the average energy consumption by 33% compared to the baselines. Moreover, it reduces the inference latency by more than 54%.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*; • **Computing methodologies** → **Intelligent agents**.

KEYWORDS

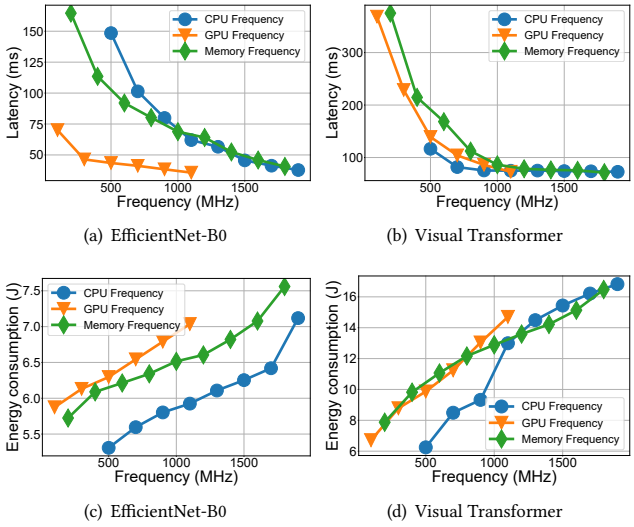
Edge Computing; Collaborative Inference; DVFS; Reinforcement Learning.

ACM Reference Format:
Ziyang Zhang, Yang Zhao, and Jie Liu. 2023. Poster Abstract: DVFO: Dynamic Voltage, Frequency and Offloading for Efficient AI on Edge Devices. In *The 22th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '23), May 9-12, 2023, San Antonio Texas, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583120.3589819

1 INTRODUCTION

In many edge intelligence applications, such as scene perception in autonomous driving, defect detection in industry and face recognition in smartphones, *etc*, edge computing devices are capable of executing deep neural networks (DNN) in real-time by combining conventional Internet of Things (IoT) infrastructure and specific hardware accelerators, *e.g.*, GPUs.

However, the DNN inference performance is limited by resources on edge devices, such as CPU, GPU and memory. In order to understand the impact of CPU, GPU and memory frequencies on inference latency and energy consumption, we conducted a set of experiments and the results are shown in Fig. 1. There are two challenges with regard to performing complex DNN inference on resource-constrained edge devices:



(a) EfficientNet-B0 (b) Visual Transformer

(c) EfficientNet-B0 (d) Visual Transformer

Figure 1: Inference latency and energy consumption with different CPU, GPU and memory frequencies for EfficientNet-B0 and Visual Transformer. Testbed: NVIDIA Xavier NX.

- **Different resources have different impacts on model performance.** We can see from Fig.1(a) and Fig.1(b) that EfficientNet-B0 is a memory-intensive DNN according to computation density (*i.e.*, the amount of computation per unit of memory access). It means that the performance bottleneck depends on the CPU and memory frequencies. On the other hand, Visual Transformer is a computation-intensive DNN, where GPU frequency dominates performance.
- **High frequency leads to high energy consumption.** From the Fig.1(c) and Fig.1(d) we can see the higher calculation frequency, the larger energy consumption. However, increasing the frequency did not reduce more latency. So we need to figure out the optimal frequency to trade-off energy consumption and latency.

Different from the single solution (*i.e.*, dynamic voltage frequency scaling (DVFS) or edge-cloud collaborative), we not only consider DVFS, but also use edge-cloud collaborative inference to minimize the overall energy consumption that includes computing and offloading energy consumption for edge devices. We propose DVFO, a deep reinforcement learning (DRL)-based DVFS technique that jointly optimizes the frequency of edge devices and the ratio of offloaded feature maps in edge-cloud collaborative inference.

IPSN '23, May 9–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0118-4/23/05…$15.00
https://doi.org/10.1145/3583120.3589819

Table 1: The computational frequency of the edge-cloud collaboration device.

	GPU	CPU Freq.	CPU Freq.	Memory Freq.
Edge	Xavier NX	1900MHz	1100MHz	1866MHz
Cloud	RTX 3080	2900MHz	1440MHz	2933MHz

2 PROBLEM DEFINITION

We consider N independent and non-preemptive tasks, denoted as $\mathcal{X} = (x_1, x_2, ..., x_N)$. The overall energy consumption E_i^{total} of the edge device for x_i consists of the computing energy consumption E_i^c and the offloading energy consumption E_i^o, *i.e.*, $E_i^{total} = E_i^c + E_i^o$. Specifically, E_i^c depends on the local computing time c_i and computing power r_i^c, denoted as $E_i^c = \sum_{i=0}^{N} c_i \cdot r_i^c$, where r_i^c is proportional to the square of the voltage V^2 and the frequency f_i, *i.e.*, $r_i^c \propto V^2 \cdot f_i$. In addition to the CPU frequency, we additionally consider GPU and memory frequencies, denoted as f_i^G, f_i^M, respectively, *i.e.*, $f_i = (f_i^C, f_i^G, f_i^M)$. E_i^o is related to the network bandwidth $\mathcal{B}$, the size of feature map to be offloaded m_i, and the offloading power of edge device r_i^o, denoted as $E_i^o = \frac{m_i \cdot r_i^o}{\mathcal{B}}$. Our objective is to minimize the energy consumption of edge devices, while satisfying system operation with a minimum frequency:

$$\begin{aligned} &min. E_i^{total} \\ &s.t. (f_{min}^C, f_{min}^G, f_{min}^M) \leq (f_i^C, f_i^G, f_i^M) \end{aligned} \tag{1}$$

We model the optimization objective in Eq. (1) as a markov decision process (MDP). We utilize DQN-based reinforcement learning algorithm to control the frequency and the ratio of offloaded feature maps for each task. Here we define the type ϕ_i of DNN x_i, and the current network bandwidth $\mathcal{B}$ as state. We set the calculation frequency f_i and the ratio of the offloaded feature map ξ for x_i as actions, and we transform Eq. (1) into a reward in DRL:

$$r = -E_i^{total} \tag{2}$$

3 SYSTEM AND PRELIMINARY EVALUATION

We designed a DVFS-based edge-cloud collaboration framework in Fig. 2. We first utilize a feature evaluator on edge device to evaluate the current DNN characteristics with negligible overhead. The feature evaluator uses computation density to evaluate whether the DNN is computation-intensive or memory-intensive workload. The DRL-based DVFS module (*i.e.*, DVFO) learns DNN characteristics and network bandwidth to optimize frequency and the ratio of offloaded feature maps for each task. After calculating the ratio of offload feature map, we need to partition the same DNN [1] that deploy to edge and cloud, respectively.

We compare DVFO with edge-only, cloud-only, and two cloud-edge collaborative methods, i.e. AppealNet [2] and DRLDO [3]. Table 1 is the computational frequency of an edge device and a cloud server. We use trickle, a lightweight bandwidth control suite to set the transmission rate of the network bandwidth to $5Mbps$. It can be observed in Fig. 3 that DVFO consistently outperforms all baselines. Specifically, the average runtime energy consumption of two DNNs using DVFO is 18%, 31%, 39%, and 43% lower than DRLDO, AppealNet, cloud-only, and edge-only, respectively. Meanwhile, DVFO reduces the average inference latency by 20%~54%.

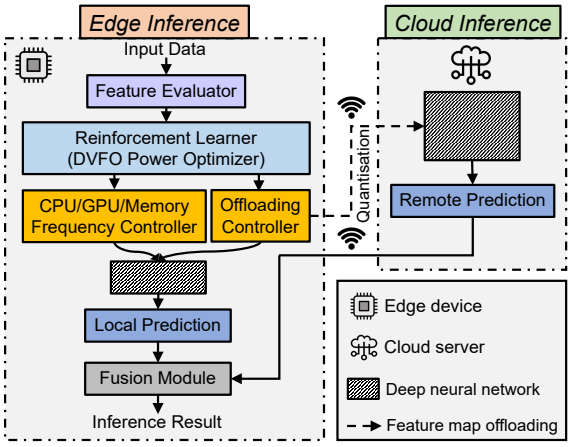


Figure 2: Overview of proposed edge-cloud collaborative inference framework.

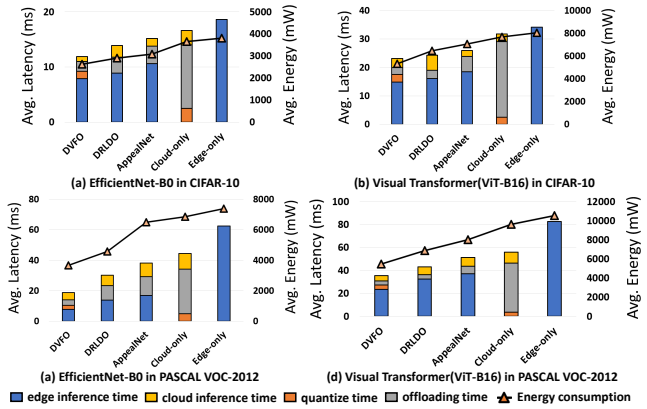


Figure 3: Comparison of inference latency and energy consumption for EfficientNet-B0 and Visual Transformer on CIFAR-100 and PASCAL VOC-2012 datasets.

4 CONCLUSION AND FUTURE WORK

This paper proposes a DRL-based DVFS technique, namely DVFO, to jointly optimize the frequency of edge device and the ratio of offloaded feature maps according to DNN characteristics and network bandwidth. Preliminary experimental results show that DVFO effectively reduces energy consumption and latency compared to baselines. In the future, we plan to explore the robustness of the system over different network bandwidths. Besides, we will explore more efficient DRL-based algorithms to reduce runtime overhead.

REFERENCES

[1] En Li, Liekang Zeng, Zhi Zhou, and Xu Chen. 2019. Edge AI: On-demand accelerating deep neural network inference via edge computing. *IEEE Transactions on Wireless Communications* 19, 1 (2019), 447–457.

[2] Min Li, Yu Li, Ye Tian, Li Jiang, and Qiang Xu. 2021. AppealNet: An Efficient and Highly-Accurate Edge/Cloud Collaborative Architecture for DNN Inference. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 409–414.

[3] Saroj Kumar Panda, Man Lin, and Ti Zhou. 2022. Energy Efficient Computation Offloading with DVFS using Deep Reinforcement Learning for Time-Critical IoT Applications in Edge Computing. *IEEE Internet of Things Journal* (2022).