Poster Abstract: E4: Energy-Efficient Early-Exit DNN Inference Framework for Edge Video Analytics

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

Deep neural networks (DNNs) are becoming extremely popular in video analytics applications at the edge. However, compute-intensive DNNs pose new challenges to achieve energy-efficient DNN inference on resource-constrained edge devices. In this paper, we propose E4, an energy-efficient DNN inference framework for edge video analytics. First, E4 analyzes video frame complexity by employing an attention-based cascade module that automatically determines DNN exit points. Second, E4's just-in-time (JIT) profiler leverages coordinate descent search to co-optimize the CPU and GPU clock frequencies for each layer before the DNN exit point. Preliminary experimental results show that E4 outperforms exiting methods in terms of power consumption and inference latency.

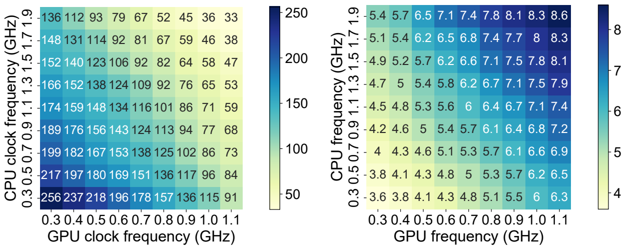
KEYWORDS

Edge Video Analytics, DVFS, Early-Exit DNN

ACM Reference Format:
Ziyang Zhang, Yang Zhao, and Jie Liu. 2023. Poster Abstract: E4: Energy-Efficient Early-Exit DNN Inference Framework for Edge Video Analytics. In Proceedings of *The 21th ACM International Conference on Embedded Networked Sensor Systems (SenSys), Nov 13-15, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3625687.3628383

1 INTRODUCTION

The continuously growing video size poses new challenges to edge video analytics. On the one hand, due to cost and volume limitations, edge devices have less computation resources than cloud servers, so that they are not equipped to match the video analytics workload where DNNs are deployed. On the other hand, DNN models are compute-intensive, requiring more power to achieve high performance. This poses a nontrivial challenge for low-power edge devices. DVFS (Dynamic Voltage and Frequency Scaling), as a popular power management technology, aims to trade off power consumption and performance by dynamically scaling CPU or GPU voltage-frequency at runtime. Nonetheless, it is still a challenge to implement a DVFS well-suited for edge video analytics to enable energy-efficient DNN inference.



(a) Inference latency (ms)

(b) Power consumption (W)

Figure 1: The effects of CPU and GPU clock frequencies on (a) inference latency (ms) and (b) power consumption (J). The x-axis represents the GPU clock frequency, and the y-axis represents the CPU clock frequency. We use Efficientnet-B0 [2] on NVIDIA Xavier NX edge GPU with 8GB DRAM.

Despite prior works have specifically customized various learning-based DVFS governors, the performance of tasks are inevitably sacrificed while reducing power consumption. As a motivating example, we use *zTT* [1], a state-of-the-art learning-based DVFS governor, on an NVIDIA Jetson Xavier NX edge device to report the impact of CPU and GPU frequencies on inference latency and power consumption during executing the EfficientNet-B0 [2] DNN model. As shown in Fig. 1, *the higher the processor clock frequency, the lower the inference latency and the higher the power consumption.* For instance, if we want to achieve 30fps video analytics (corresponding to 30ms inference latency), the CPU and GPU clock frequencies have to be adjusted to the highest level (corresponding to 1.9GHz and 1.1GHz, respectively), the power consumption, unfortunately, increases dramatically to 8.6W. It reveals that DVFS does not enable low-latency and less power consumption at the same time, which motivates us to design an efficient power management technology.

In this paper, we introduce E4, an energy-efficient DNN inference framework for edge video analytics that integrates early-exit (*i.e.*, a mechanism that enables early-exit at different points during DNN inference), which can adapt to video frame complexity and DNN model diversity.

2 SYSTEM AND PRELIMINARY EVALUATION

E4 tackles the problem of energy-efficient DNN inference for edge video analytics. Fig. 2 depicts the overview of E4, which consists of two phases: (1) *Offline Phases* has two components: (i) *Accumulated Feature Pooling* and (ii) *Attention-Based Early-Exit*, which are responsible for analyzing video frame complexity and determining the DNN exit point respectively, as input to *Just-In-Time Profiler* in

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3628383

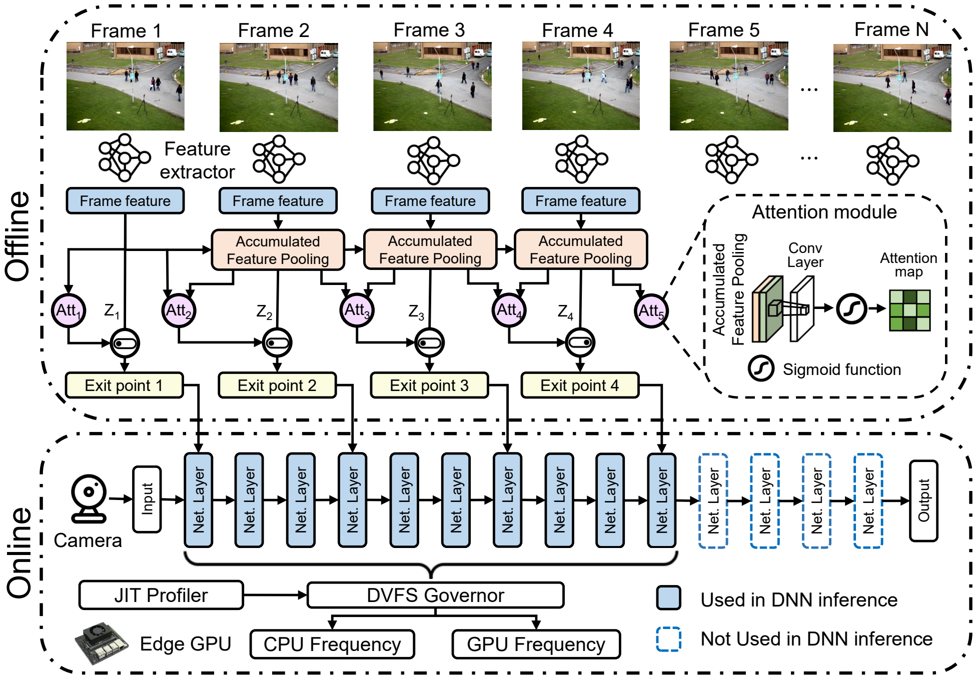


Figure 2: The overview of E4.

online phase. (2) *Online Phase* integrates two components: (i) *Just-In-Time Profiler* and (ii) *DVFS Governor*. The former is responsible for dynamically co-optimizing the CPU and GPU frequencies of each layer before the DNN exit point, while the latter is used for dynamic frequency scaling. We have implemented E4 using Python 3.6 on two NVIDIA Jetson series of edge devices. EfficientNet-B0 [2] is placed five DNN exit points and is pretrained on ImageNet dataset. Furthermore, we use the ActivityNet-v1.3 video analytics dataset.

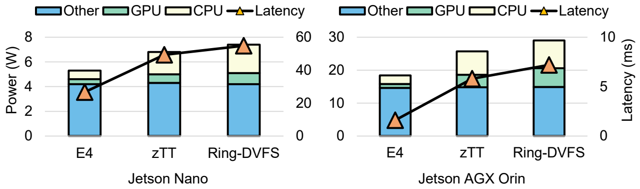


Figure 3: Power consumption and inference latency comparison of E4 against other learning-based DVFS methods.

We compare E4 with two learning-based DVFS methods, *i.e.*, *zTT* [1] and *Ring-DVFS* [3]. Since *Ring-DVFS* [3] is designed only for embedded devices with CPUs, we extend it to edge devices with CPU-GPU heterogeneous processors for fair comparison. We report the performance of all methods in terms of power consumption and inference latency. As shown in Fig. 3(a), we first examined the video analytics task using EfficientNet-B0. Unsurprisingly, E4 consistently outperforms *zTT* [1] and *Ring-DVFS* [3]. Overall, E4 enhances energy-efficient compared to baselines, achieving 20%~37% energy-saving and 1.3×~2.2× inference speedup, respectively. Intuitively, compared to *zTT* [3] and *Ring-DVFS* [3] which only unilaterally optimize the clock frequency of heterogeneous processors, the performance improvement of E4 is attributed to DNN's early-exit mechanism, which significantly reduces computation cost and power consumption by partial DNN inference. In addition, we find that compared with edge devices with low computility, E4 brings more significant performance improvement to edge devices with high computility. For instance, the performance improvement of Jetson AGX Orin, which has the highest computility, is 37% higher on average than that of Jetson Nano with the lowest computility. The results are attributed to the fact that high computility means a larger frequency range, so that E4 has a larger optimization space.

3 CONCLUSION AND FUTURE WORK

This paper proposes E4, an energy-efficient DNN inference framework for edge video analytics. E4 introduces two design knobs to enable energy-efficient DNN inference: DNN's early-exit mechanism and a novel DVFS governor, which are responsible for determining the DNN exit point as well as optimal CPU and GPU clock frequencies, respectively. Preliminary experimental results show that E4 outperforms exiting methods in terms of power consumption and inference latency. For the future work, our proposed E4 can be further combined with various existing model compression techniques to achieve higher energy-efficient DNN inference.

REFERENCES

[1] Seyeon Kim, Kyungmin Bin, Sangtae Ha, Kyunghan Lee, and Song Chong. 2021. zTT: Learning-based DVFs with zero thermal throttling for mobile devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 41–53.

[2] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[3] Amir Yeganeh-Khaksar, Mohsen Ansari, Sepideh Safari, Sina Yari-Karin, and Alireza Ejlali. 2020. Ring-DVFS: Reliability-aware reinforcement learning-based DVFS for real-time embedded systems. *IEEE Embedded Systems Letters* 13, 3 (2020), 146–149.