MULTIUSER EDGE INTELLIGENCE ENERGY-MANAGEMENT NEURALNET (NN) MAXIMIZING TASK COMPLETION RATE WITH PARTITIONING AND OFFLOADING

Suman B[1], Dr. Pramod Pandurang Jadhav[2]

Article History: Received: 02.11.2022 **Revised:** 18.11.2022 **Accepted:** 23.12.2022

Abstract:

NNs have emerged as a critical technology for edge intelligence. It is not possible to run large-scale NNs directly on Internet of Things (IoT) devices with limited energy since doing so requires a lot of resources and energy. By outsourcing certain NN layers to run on the edge server, NN partitioning offers a workable solution to this issue. Edge servers' resources, however, are often constrained. In such a realistic setting, it results in an resource and energy optimization-constrained problem.. Due to this, we look at an intractable nonlinear optimization issue known as NNoffloading and partitioning over a multiuser resource-constrained setting. We divide the issue into two smaller issues and provide an Energy-Management Neural Net Partitioning and Offloading (EMNPO) technique using dynamic programming and the theorem of minimum cut/maximum flow to find the solution in polynomial time. Lastly, we investigate how the energy limitation, NN type, and device count affect the effectiveness of the EMNPO. The technique suggested could dramatically improve the NN inference task's completion rate in comparison to other approaches as demonstrated by the simulation results of real NN models.

[1,2]Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University, Indore (M.P.) – 452010

DOI: 10.53555/ecb/2023.12.3.251

1. Introduction

With the integration of wireless technologies, microelectromechanical systems, and the internet, the Internet of Things has evolved. It refers to the usage of intelligently linked devices to gather data from embedded sensors and other physically present items that are made possible via the internet. They have a significant impact on customers, the community, and the issues, allowing life-improving services in a wide range of applications that fall under the categories of intelligent living, environment, and business. Many options exist for the IOT to improve service quality. technique using dynamic programming and the theorem of cut/maximum flow to answer it for polynomial-time infrastructure, platforms, and other designed services are situated remotely and process data in a distant location. So, it takes more time and energy to transmit, analyse, and retrieve data, which makes timely retrieval less effective. The delays caused by by jobs that must be completed by a certain time may cause deaths or any other illicit access that changes the information. So, in order to have timely access, it becomes necessary to create a network adjacent to the devices or sensors.

Several fields, like face recognition and self-driving cars, have seen considerable success using deep neural networks (NNs) [2], thanks to the invention of deep learning [1]. NN may be used on intelligent Internet of Things (IoT) devices for completing inference tasks, such as wearable technology and mobile phones [4], thanks in part to the burgeoning IoT sector [3]. Large-scale NNs may often provide more accurate analytical findings, as has been shown [5]. Deeper network architectures need more processing resources and energy, which severely restricts the use on energy-constrained IoT devices of large-scale NNs. We examine an original energy-aware NN inference by simultaneously analysing NN Model partitioning and allocation of resources to address NN for inference tasks. issue , with the intention to increase the number of energetically aware NN inference tasks that are completed.

2. Related Works

The paper goes on to discuss related work that examines the efficacy of cloud computing, as suggested by Berl and al. [1], suggest an alternative approach to energy efficiency on the operation of systems and networks in the allocation of resources, presents a survey on the effective ways that cloud computing can cut down on energy use, and discusses the field's future directions. To improve the network's efficiency in terms of service quality, Marinos, et al. [2] discuss the advantages of explaining the benefits of defining the cloud community. While cloud computing appears to be a viable option, Moreno et alresearch . shows that it faces additional challenges when it comes to service delivery. [3] study of the issues that cloud computing will bring to the Internet of things. Tan et alintroduction . The internet of things [4] provides an overview of the many applications that the technology is used for, creating connections between the surrounding physical objects with the aid of software that is both integrated into them and linked to them through the internet. Choi et al. [6] demonstrated the allocation of resources to manage the data that is collected. They also showed the efficiency of resource allocation by cloud computing to internet of things. Additionally, Cloud Edge network, which was created to handle latency-sensitive real-time applications, was surveyed by Mahmud et al. [9] in order to shorten access times and have timely access to information. The Cloud Edge network was developed to be closer to the user in order to shorten access times and avoid gaps that could allow for unauthorised access. As a challenge to cloud computing, According to Zanafi et al. [10], the Cloud Edge network offers sustainable smart environments by lowering service energy consumption. The resource allocation provided by Zhang, et al. [11] increases resource consumption but leads in higher costs. Another approach presented by Lan, et al. [12] is the use of cloud edge networks. Cloud Edge network for resource allocation and offloading D2D assisted Cloud Edge computing networks. More resource utilisation results are offered, however the energy usage for resource allocation does not improve. So, the recommended approach attempts to provide the optimal resource use while reducing energy consumption for the IOT-based sensed data.

3. System Model

our system consists of one Access Point (AP) As shown in Figure 1, connected to N IoT devices as well as a colocated edge server. The transmission delay between the edge server and AP is disregarded since they are connected through a wire. Also, the edge server has a limited amount of computational capacity; each device has a

single NN inference job with a particular NN model and a modest power need (Condition for task I's end). Remember that the models are locally edge-stored and pretrained. Local devices use less energy by offloading certain NN layers to operate on the edge server.

NN deduction. NN inference uses a lot less CPU and GPU resources than NN training. Consequently, We look at edge servers based on CPUs, where each CPU core corresponds to a thread, as an alternative to deploying expensive hardware accelerators (such as GPUs) for NN inference. The majority of contemporary NN models typically consist of multiple crucial components called "layers." The identity of a "layer" denotes a collection of computations directed towards a certain set of inputs with a comparable amount of effort. Each layer of the model accepts the output of the layer before it as an input, as the roughly GoogleNet model seen in Figure 2. In this article, layers are seen as the fundamental components of a NN model. We designate NN as a Directed Acyclic Graph to match the structure of a genuine neural network (DAG).

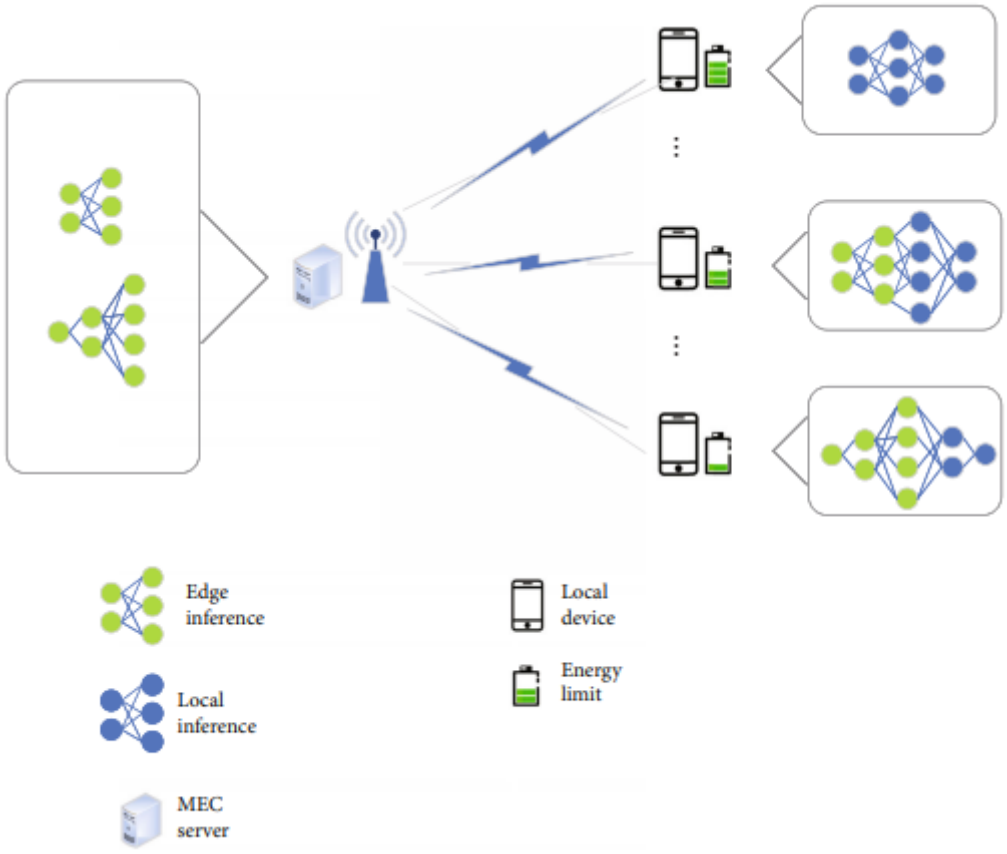


Figure 1: Energy-Management NN inference system model with numerous IoT devices.

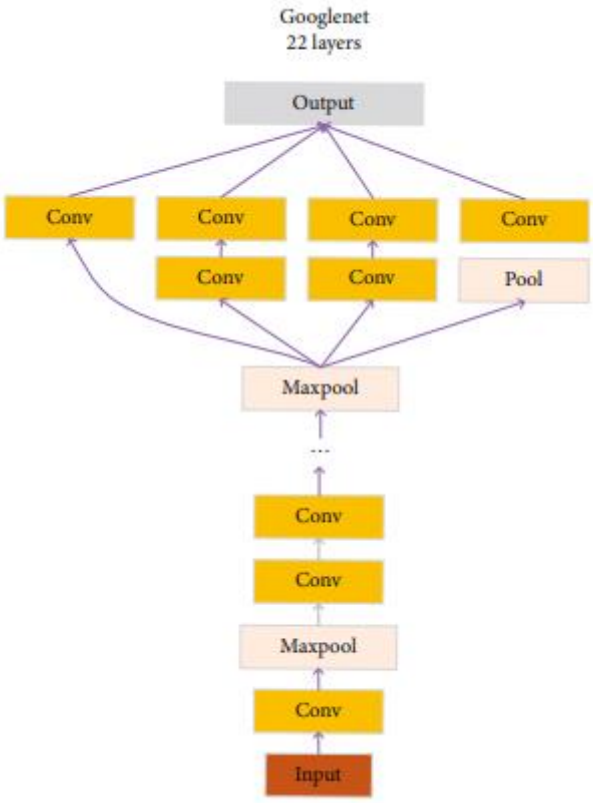


Figure 2: The general network architecture of GoogleNet.

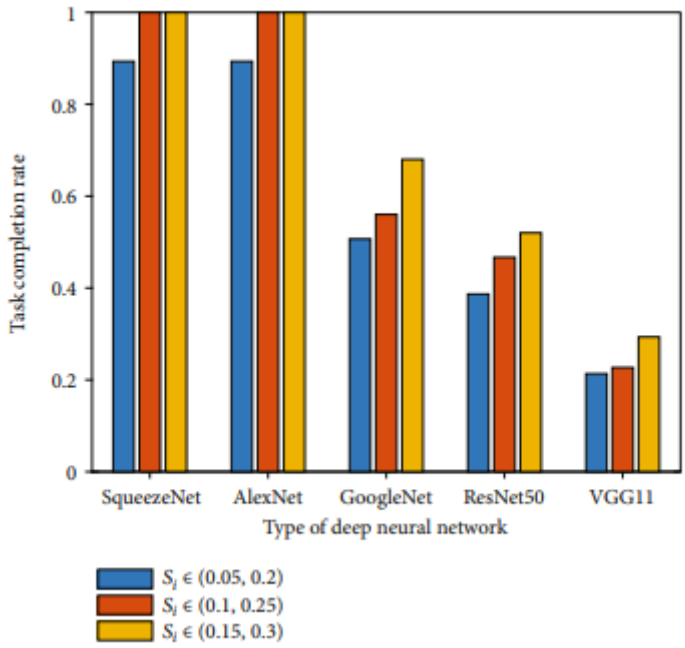
The energy consumption of the tasks we took into consideration specifically pertains to IoT device energy consumption. It is important to note that while offloaded layers are running at the edge, we also take the device's idle power consumption into account. We disregard the energy used to get the provided findings due to the little quantity of data returned.

4. Proposed Energy-Management NN Partitioning and Offloading Model Results

We propose the Energy-Management NN Partition and Offloading (EMNPO) method as a solution to the issue raised in Section 2. This algorithm adaptively decides how to divide and offload NN based on the task's low power requirements and the edge server's resource constraints.

We offer three heuristic algorithms as comparison standards to assess the effectiveness of EMNPO in terms of completion rate. All jobs are allowed to be offloaded by EDGY-GREEDY, and the Greedy approach chooses which tasks to offload based on the server resources available. With limited server resources, RANDOF selects a few jobs at random for offloading. LOCAL-ONLY causes all operations to be carried out locally. Moreover, we contrast the efficiency of energy usage between EMNPO and Neurosurgeon, a cutting-edge NN partitioning approach. Moreover, the completion rate of AlexNet is higher than that of ResNet and GoogleNet (32% vs. 45.4%, respectively). They need more FLOPs than AlexNet (14.53G and 62.77G). Naturally, they need additional server resources in order to comply with the energy restriction.

comparing the performance of energy consumption between EMNPO and neurosurgeon. Then, using several NN types, we assessed how well EMNPO and Neurosurgeon saved energy. The results establish the Edge-Only method as the normal procedure (the performance is normalized to the Edge-Only strategy). Figure 3 illustrates that the efficiency of energy conservation for EMNPO as well as Neurosurgeon for the topology models of the chain is similar. We can see that EMNPO performs substantially better than Neurosurgeon for the DAG topology. As compared to Neurosurgeon, EMNPO provides a 30% energy savings for DAG topology models. Next, using various edge computing capacities, we assess GoogleNet's energy usage performance. As shown in Figure 3, EMNPO consistently uses less energy than Neurosurgeon even if their energy requirements are decreasing as edge computing capability rises. This finding supports the efficacy of EMNPO.



S_j = Eneergy source in joules

Figure 3: Energy restrictions' effects on EMNPO performance.

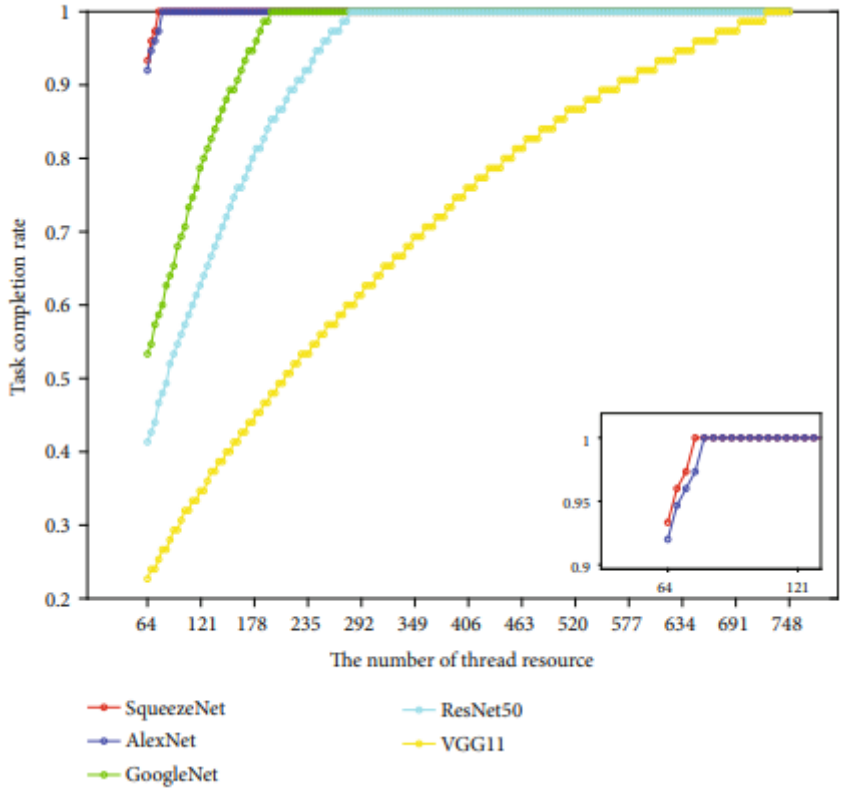


Figure 4: Variations in the total energy used depending on the pace of completion for various NN kinds.

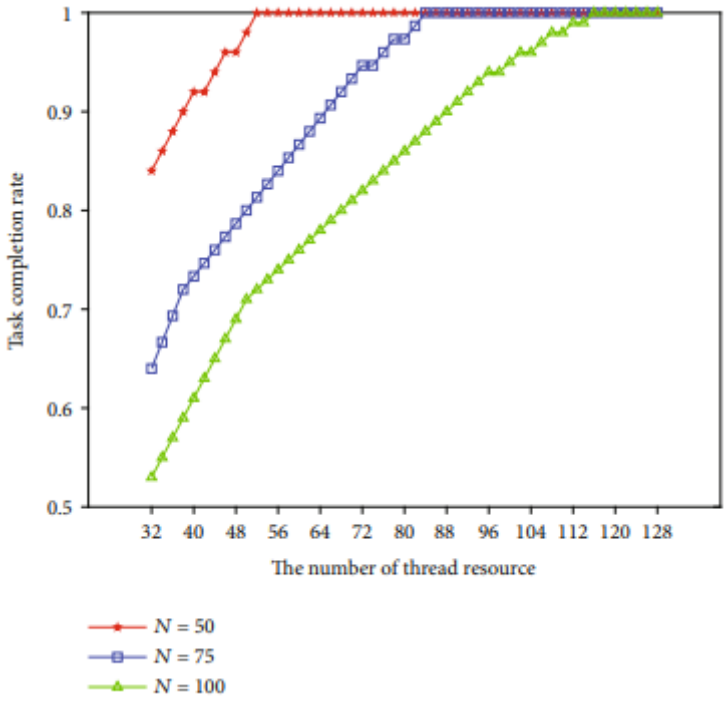


Figure 5: Energy fluctuation while achieving completion rate 1 with various device counts.

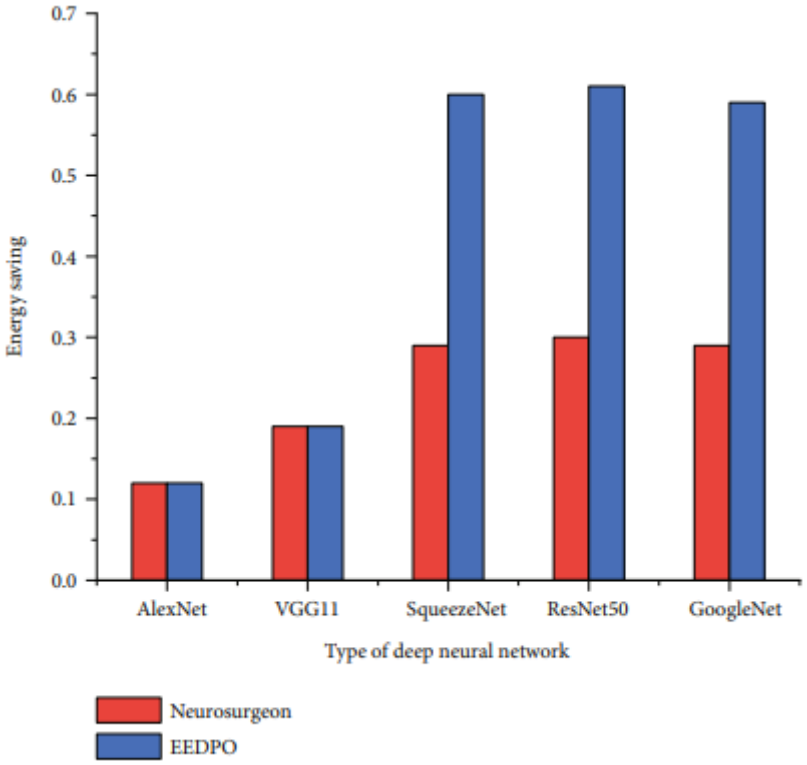


Figure 6 : Neurosurgeon and EMNPO both reduced energy consumption utilising various NN types, with the Edge-Only technique serving as the baseline.

5. Conclusion

In this paper, we tackle the problem of maximising the pace at which NN inference tasks are completed in a multiuser edge environment while also taking into consideration the resource and energy limitations of the edge server and the IoT devices. We presented an Energy-Management NN Partitioning and Offloading (EMNPO) technique to decrease task energy usage and increase server resource utilisation. The suggested algorithms have shown promise in experimental findings. Also, we describe some important findings from the assessment results (Figure 4 allows us to make an educated guess on resources to install on the edge server if the number of devices hits a certain threshold.). The adaption of the suggested partition technique will be improved by taking into account more complex situations involving environmental variation, such as network latency, wireless channel condition, and server failure. In addition, we want to research the issue of NN offloading and inference in wireless-powered MEC systems.

6. References

1. Berl, Andreas, Erol Gelenbe, Marco Di Girolamo, Giovanni Giuliani, Hermann De Meer, Minh Quan Dang, and Kostas Pentikousis. "Energy-efficient cloud computing." The computer journal 53, no. 7 (2010): 1045-1051.
2. Marinos, Alexandros, and Gerard Briscoe. "Community cloud computing." In IEEE International Conference on Cloud Computing, pp. 472-484. Springer, Berlin, Heidelberg, 2009.
3. Moreno-Vozmediano, Rafael, Rubén S. Montero, and Ignacio M. Llorente. "Key challenges in cloud computing:

Enabling the future internet of services." IEEE Internet Computing 17, no. 4 (2012): 18-25.

4. Tan, Lu, and Neng Wang. "Future internet: The internet of things." In 2010 3rd international conference on advanced computer theory and engineering (ICACTE), vol. 5, pp. V5-376. IEEE, 2010.
5. Chien, Wei-Che, Chin-Feng Lai, M. Shamim Hossain, and Ghulam Muhammad. "Heterogeneous Space and Terrestrial Integrated Networks for IoT: Architecture and Challenges." IEEE Network 33, no. 1 (2019): 15-21.
6. Choi, Yeongho, and Yujin Lim. "Optimization approach for resource allocation on cloud computing for iot." International Journal of Distributed Sensor Networks 12, no. 3 (2016): 3479247.
7. Zhou, Zhenyu, Mianxiong Dong, Kaoru Ota, Guojun Wang, and Laurence T. Yang. "Energy-efficient resource allocation for D2D communications underlaying cloud-RAN-based LTE-A networks." IEEE Internet of Things Journal 3, no. 3 (2015): 428-438.
8. Mishra, Bhabani Shankar Prasad, Himansu Das, Satchidananda Dehuri, and Alok Kumar Jagadev, eds. Cloud Computing for Optimization: Foundations, Applications, and Challenges. Springer International Publishing, 2018.
9. Mahmud, Redowan, Ramamohanarao Kotagiri, and Rajkumar Buyya. "Fog computing: A taxonomy, survey and future directions." In Internet of everything, pp. 103-130. Springer, Singapore, 2018.
10. Zanafi, Sarah, Noura Aknin, Maurizio Giacobbe, Marco Scarpa, and Antonio Puliafito. "Enabling Sustainable Smart Environments Using Fog Computing." In 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1-6. IEEE, 2018.
11. Zhang, Wenyu, Zhenjiang Zhang, and Han-Chieh Chao. "Cooperative fog computing for dealing with big data in the internet of vehicles: Architecture and hierarchical resource management." IEEE Communications Magazine 55, no. 12 (2017): 60-67.
12. Lan, Yanwen, Xiaoxiang Wang, Dongyu Wang, and Zhaolin Liu. "Task Caching, Offloading and Resource Allocation in D2D-Aided Fog Computing Networks." IEEE Access (2019).
13. X.Tang,X.Chen,L.Zeng,S.Yu,andL.Chen,"Jointmultiuser DNN partitioning and computational resource allocation for collaborative edge intelligence," IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9511–9522, 2021.
14. Z. Xu, L. Zhao, W. Liang et al., "Energy-aware inference offloading for DNN-driven applications in mobile edge clouds," IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 4, pp. 799–814, 2021.
15. M.Xue,H.Wu,G.Peng,andK.Wolter," DDPQN:anefficient DNN offloading strategy in local-edge-cloud collaborative environments," IEEE Transactions on Services Computing, vol. 15, no. 2, pp. 640–655, 2022.
16. M. Xue, H. Wu, R. Li, M. Xu, and P. Jiao, "EosDNN: an efficient offloading scheme for DNN inference acceleration in local-edge-cloud collaborative environments," IEEE Transactions on Green Communications and Networking, vol. 6, no. 1, pp. 248–264, 2022.
17. H. Wu, W. J. Knottenbelt, and K. Wolter, "An efficient application

partitioning algorithm in mobile environments," IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 7, pp. 1464–1480, 2019.

18. X. Chen, M. Li, H. Zhong, Y. Ma, and C.-H. Hsu, "DNNOff: offloading DNN-based intelligent IOT applications in mobile edgecomputing,"IEEETransactionso nIndustrialInformatics, vol. 18, no. 4, pp. 2820–2829, 2022.
19. J. Li, W. Liang, Y. Li, Z. Xu, and X. Jia, "Delay-aware DNN inference throughput maximization in edge computing via jointly exploring partitioning and parallelism," in 2021 IEEE 46th Conference on Local Computer Networks (LCN), pp. 193–200, Edmonton, AB, Canada, October 2021.
20. Z.Gong,H.Ji,C.W.Fletcher,C.J.Hugh es,S.Baghsorkhi,and J. Torrellas, "Save: sparsity-aware vector engine for accelerating DNN training and inference on CPUs," in 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 796–810, Athens, Greece, October 2020.
21. A. V. Nori, R. Bera, S. Balachandran et al., "Proximu $: efficiently scaling dnn inference in multi-core CPUs through near-cache compute," 2020, https://arxiv.org/abs/2011.11695.
22. Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 9, pp. 1124–1137, 2004.