IEEE Hong Kong 6G Wireless Summit
11-12 September 2024
Hong Kong, China

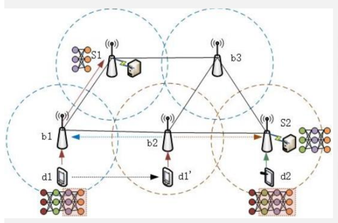
Mobility and Cost Aware Inference Accelerating Algorithm for Edge Intelligence

Xin Yuan, Ning Li, KangWei, Wenchao Xu, Quan Chen, HaoChen, Song Guo

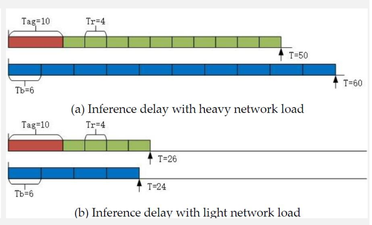
Abstract The edge intelligence(EI) has been widely applied recently. Splitting the model between device, edge server, and cloud can significantly improve the performance of EI. The model segmentation without user mobility has been investigated in detail in previous studies. However, in most EI use cases, the end devices are mobile. Few studies have been conducted on this topic. These works still have many issues, such as ignoring the energy consumption of mobile device, inappropriate network assumption, and low effectiveness on adapting user mobility, etc. Therefore, to address the disadvantages of model segmentation and resource allocation in previous studies, we propose mobility and cost aware model segmentation and resource allocation algorithm for accelerating the inference at edge(MCSA). Specifically, in the scenario without user mobility, the loop iteration gradient descent (Li-GD) algorithm is provided. When the mobile user has a large model inference task that needs to be calculated, it will take the energy consumption of mobile user, the communication and computing resource renting cost, and the inference delay into account to find the optimal model segmentation and resource allocation strategy. In the scenario with user mobility, the mobility aware Li-GD (MLi-GD) algorithm is proposed to calculate the optimal strategy. Then, the properties of the proposed algorithms are investigated, including convergence, complexity, and approximation ratio. The experimental results demonstrate the effectiveness of the proposed algorithms.

Network model and problem statement

Based on the network model and the proposed models (the ***inference delay, energy consumption model, and resource renting cost models***), the problems that will be solved in this study (finding the minimum energy consumption, minimum inference delay, and mini-mum resource utilization simultaneously) are described in detail.



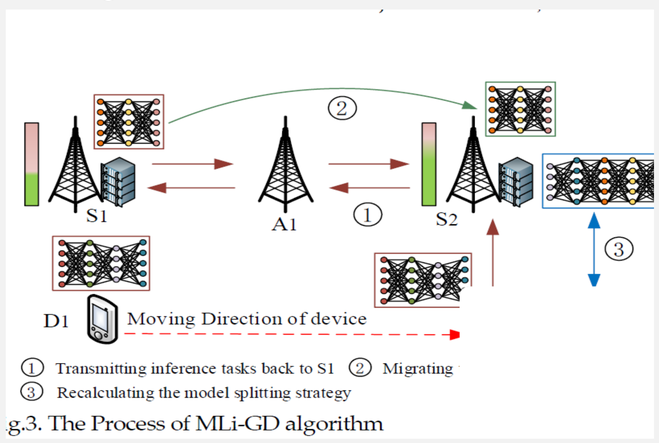
Network model



Inference delay under different network load

Optimal model segmentation and resource allocation

Under the scenario in which the end devices are motionless or can move within the coverage of AP, we investigate the optimal solution for Po which is. We propose an approximation algorithm—Li-GD algorithm—for Po, the properties of Li-GD algorithm, including its convergence and complexity are investigated. ***For the scenario in which users can move*** (as demonstrated in figure below), calculating the optimal solution becomes difficult. The mobility aware Li-GD (MLi-GD) algorithm is proposed and conclusions are as follows: the utility function is differentiable; the approximation ratio of MLi-GD algorithm is ε; the complexity of the MLi-GD algorithm is the same as that of the Li-GD algorithm; the MLi-GD algorithm is convergent .

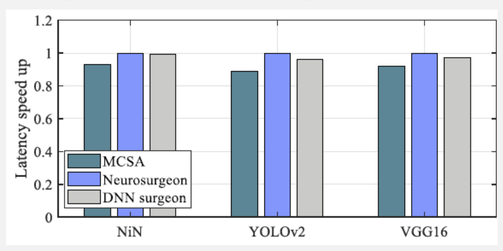


g.3. The Process of MLi-GD algorithm

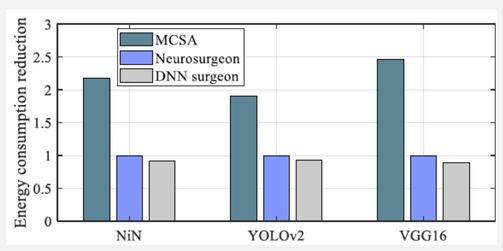
The Process of MLi-GD algorithm

Performance evaluation without user mobility

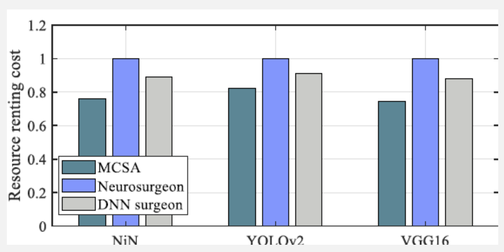
We compare the performance of MCSA with Device-Only, Edge-Only, Neurosurgeon, and DNN surgeon, the performance is normalized to the Device-Only method. The performance of MCSA is compared with Neurosurgeon and DNN surgeon. The latency speedup in MCSA and DNN surgeon is similar but a little better than that in Neurosurgeon. The energy consumption reduction in MCSA is much better than that in Neuro-surgeon and DNN surgeon, and the energy consumption reduction in Neurosurgeon and DNN surgeon is similar. We can conclude that the performance of resource renting cost in MCSA is much better than that in Neuro-surgeon and DNN surgeon.



Latency speedup for different DNN models



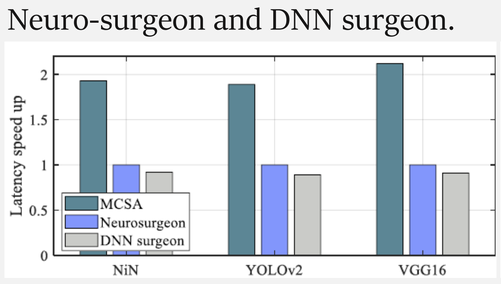
Energy consumption reduction for different DNN models



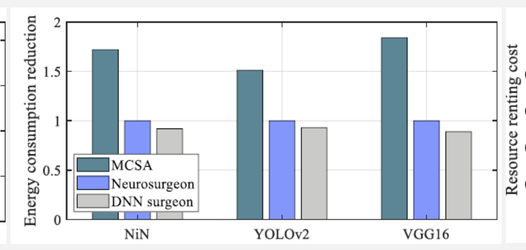
Resource renting cost for different DNN models

Performance evaluation with user mobility

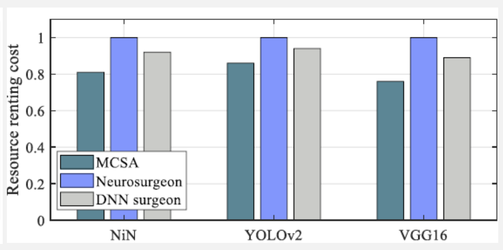
Under the scenario in which the users are mobile, the latency speedup in MCSA and DNN surgeon is similar but a little better than that in Neurosurgeon. The energy consumption reduction in MCSA is much better than that in Neuro-surgeon and DNN surgeon, and the energy consumption reduction in Neurosurgeon and DNN surgeon is similar. We can conclude that the performance of resource renting cost in MCSA is much better than that in Neuro-surgeon and DNN surgeon.



Latency speedup for different DNN models



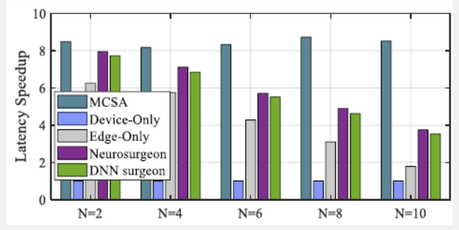
Energy consumption reduction for different DNN models



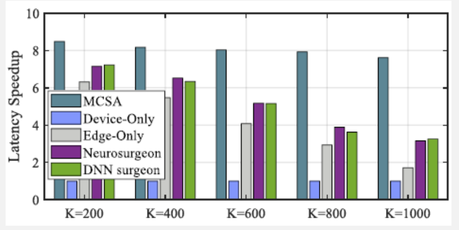
Resource renting cost for different DNN models

Performance evaluation under different network condition

We compare the performance of MCSA with Device-Only, Edge-Only, Neurosurgeon, and DNN surgeon ***under different network conditions, including different number of hops for intermediate data transmission and inference task computing load***. We can conclude that with the increasing of the number of hops, i.e., the distance between the mobile devices to its original edge server, the latency in Edge-Only, Neurosurgeon, and DNN surgeon rises. Also, with the increasing of the computing load, the performance of latency speedup of all the algorithms except for the Device-Only approach.



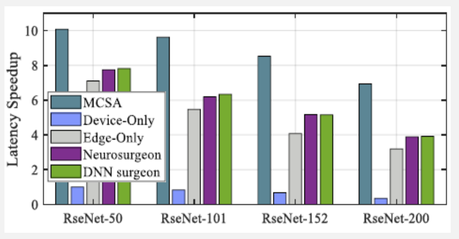
Latency speedup under different number of hops



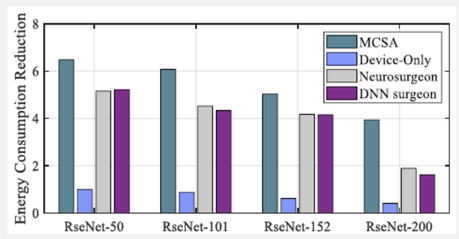
Latency speedup under different computing load

Performance evaluation under different model sizes

We compare the performance of MCSA with Device-Only, Edge-Only, Neurosurgeon, and DNN surgeon ***under different size of models, i.e., RseNet-50, RseNet-101, RseNet-152, and RseNet-200***, which are much larger than those of NiN, VGG-16, and YOLOv2. We can find that with the increasing of model size, the latency in both these five algorithms increases. And, due to the increasing of the model size, the energy consumption of both these four algorithms increases.



Latency speedup under different size of models



Energy consumption reduction under different size of models

ACKNOWLEDGMENT

This work was supported in part by the Research Grants Council of Hong Kong under the Areas of Excellence scheme grant AoE/E-601/22-R.