HAMS: Hardware-Aware Model Scheduling on Heterogeneous Platforms

Haofeng Kou - Baidu Research Institute *
Yongtao Yao - Wayne State University *
Sidi Lu - Wayne State University
Yueqiang Cheng - Baidu Research Institute
Weijia Shang - Santa Clara University
Weisong Shi - Wayne State University
Problems

How to concurrently & efficiently deploy and execute the collaborative models on heterogeneous devices with different deployment constraints?

- The real-world applications usually require collaboration of multiple DNN models on edge computing platforms to finish complicated tasks with outstanding performance
- Explosive growth in model size, computational requirements, increasing number of involved models and devices
Previous Work

One-to-One: One DNN architecture to one hardware platform

- Design a network architecture that is both accurate and efficient on a given edge device
- Train a separate model for each device of interest and each latency budget of interest
- Too resource demanding for the case-by-case deployment environment
- Not practical enough when the real-world application requires the involvement of multi-models and diverse devices at the same time
Our Research - Innovation

Many-to-Many: providing actionable insights on scheduling the efficient deployment of a group of collaborative DNN models among heterogeneous hardware devices and assessment of our proposed partition and scheduling algorithm

- The multiple models scheduling problem for the edge computing tasks in the heterogeneous environment has **not been deeply studied** yet.
- Our proposed framework is the pioneer that points out the importance of this **new research direction** with useful insights for related research.
Many-to-Many: providing actionable insights on scheduling the efficient deployment of a group of collaborative DNN models among heterogeneous hardware devices and assessment of our proposed partition and scheduling algorithm

- We have demonstrated the applicability of the proposed scheduling algorithms MFS and HFS, in three typical application scenarios of the computer vision field, with the ability of hardware adaptive self-learning to automatically schedule the deployment and execution of multiple models on heterogeneous edge services
Our Research - Result

Many-to-Many: providing actionable insights on scheduling the efficient deployment of a group of collaborative DNNs among heterogeneous hardware devices and assessment of our proposed partition and scheduling algorithm

- Our analysis reveals that HAMS can balance computation resource utilization and reduce the inference time of the whole group of models up to 28.77%.
NCO & NCA

HAMS contains two core components:

NCO - Neural Computing Optimizer responsible for training, optimizing, and transforming DNN models into a hardware-specific format so that the model can fit a given hardware platform well

NCA - Neural Computing Accelerator integrate of HAMS that contains our proposed design
Matrix Generation:
- Calculate FPS of each model running independently on each device
- Overall inference speed dependent on where the slowest speed is
Target at finding an appropriate model for edge devices

- ModelAllocations
- QueryWorstCaseModel
- QueryModel
Aim to find a suitable edge device for specific models

- DeviceAllocations
- QueryWorstCaseDevice
- QueryDevice
Single Service

The individual service models assigned to their most suitable edge devices
Overall FPS for each service will be calculated separately

<table>
<thead>
<tr>
<th>Service</th>
<th>DNN Model</th>
<th>MFS</th>
<th>HFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Face Detection</td>
<td>GPU</td>
<td>GPU</td>
</tr>
<tr>
<td></td>
<td>Age/Gender Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td></td>
<td>Emotions Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td></td>
<td>Facial Landmarks</td>
<td>CPU</td>
<td>CPU</td>
</tr>
<tr>
<td></td>
<td>Head Pose Estimation</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>P</td>
<td>Person Detection</td>
<td>GPU</td>
<td>GPU</td>
</tr>
<tr>
<td></td>
<td>Person Attributes Recognition</td>
<td>VPU</td>
<td>CPU</td>
</tr>
<tr>
<td></td>
<td>Person Reidentification</td>
<td>CPU</td>
<td>VPU</td>
</tr>
<tr>
<td>V</td>
<td>Vehicle License Plate Detection</td>
<td>GPU</td>
<td>VPU</td>
</tr>
<tr>
<td></td>
<td>Vehicle Attributes Recognition</td>
<td>VPU</td>
<td>CPU</td>
</tr>
<tr>
<td></td>
<td>License Plate Recognition</td>
<td>CPU</td>
<td>GPU</td>
</tr>
</tbody>
</table>

- **Service F**: MFS & HFS leads to the same FPS (5.64), 28.77% higher than default FPS (4.38)
- **Service P and Service V**: HAMS improve FPS by 2.58%
Multiple Service

Three sets of 11 models assigned to their most suitable edge devices
VPUs can be expanded - one model to one edge device
Overall FPS for all services & models are calculated together

<table>
<thead>
<tr>
<th>Model Name</th>
<th>MFS</th>
<th>HFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face Detection</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Age/Gender Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Emotions Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Facial Landmarks</td>
<td>VPU</td>
<td>CPU</td>
</tr>
<tr>
<td>Head Pose Estimation</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Person Detection</td>
<td>GPU</td>
<td>GPU</td>
</tr>
<tr>
<td>Person Attributes Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Person Reidentification</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Vehicle License Plate Detection</td>
<td>CPU</td>
<td>VPU</td>
</tr>
<tr>
<td>Vehicle Attributes Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
<tr>
<td>License Plate Recognition</td>
<td>VPU</td>
<td>VPU</td>
</tr>
</tbody>
</table>

- Service F/P/V shows better FPS than default FPS scheduling
Open Discussion

- **Task-Level Scheduling on Heterogeneous Platforms**
  - StarPU on HPC
  - ESTS on HCS
  - OmpSs
  - AlEbrahim

- **Neural Architecture Search**
  - MnasNet
  - DARTS - Differentiable ARchiTecture Search
  - FBNets - Facebook-Berkeley-Nets
  - Once-for-All

- **Gap between Previous Work**
  - Compared with Task-Level Scheduling
  - Compared with Neural Architecture Search
Summary

- Prove the importance of model scheduling for multiple DNNs and heterogeneous edge devices with diverse computation resources
- Key concept is *Worst-Case-First* for hardware-aware models scheduling
- Introduce and discuss two scheduling algorithms and get the evaluation results of three DNN groups on CPU, GPU and multiple VPUs
- The evaluation results demonstrate the effectiveness of HAMS on accelerating the co-inference of multi-models on the heterogeneous edge devices by up to 28.77%
Acknowledgement & QA

- Thanks for the collaboration from WSU, SCU and BRI!
- Thanks SEC20 offering the chance!
- We can be reached at: BRI & WSU & SCU
  - kouhaofeng@baidu.com
  - yongtaoyao@wayne.edu