Dynamic DNNs Meet Runtime Resource Management for Efficient Heterogeneous Computing

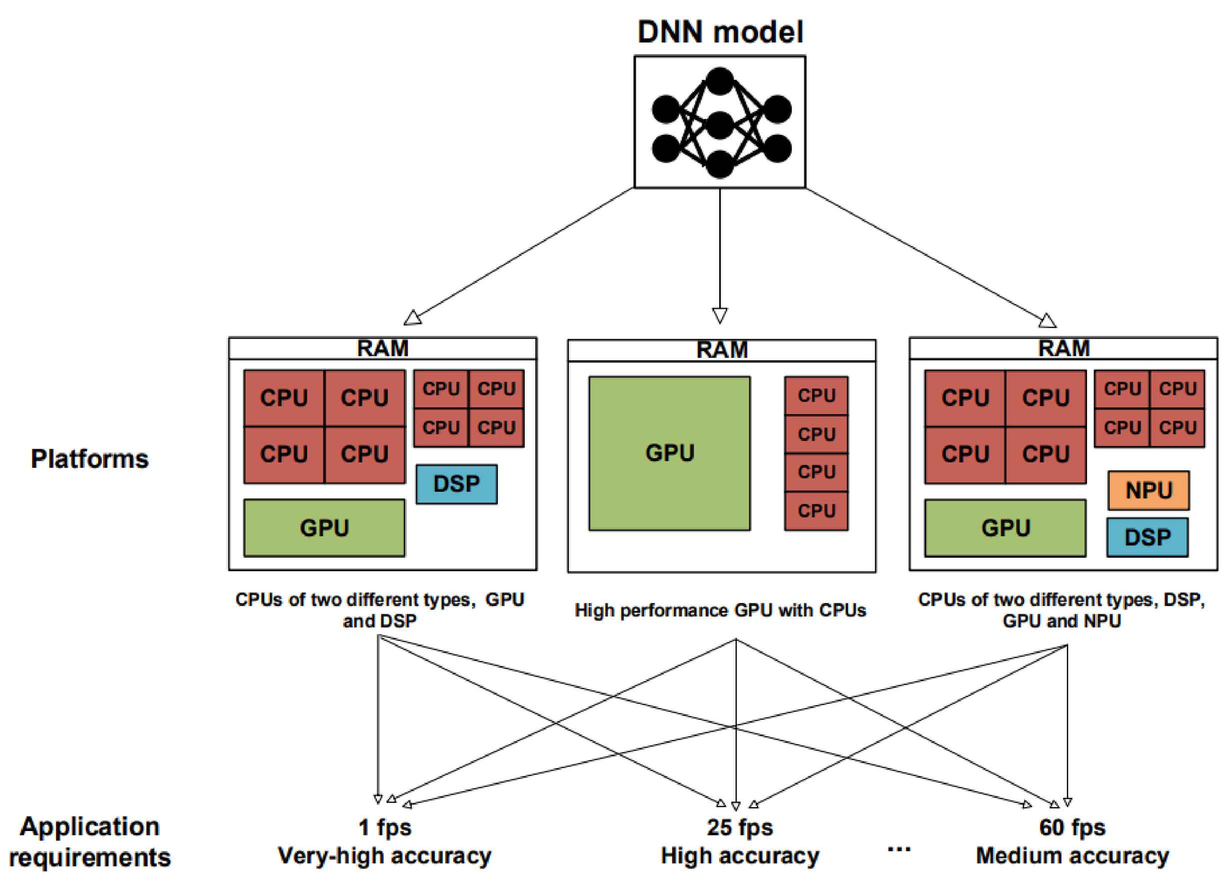
Lei Xun, Jonathon Hare, Geoff Merrett

Email: l.xun@soton.ac.uk

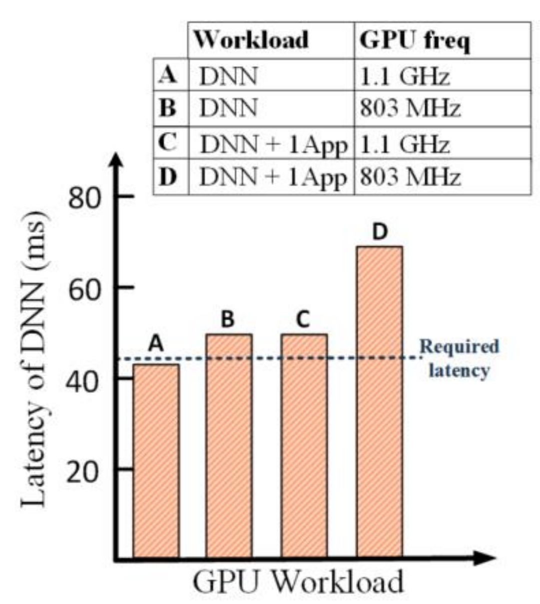
UNIVERSITY OF Southampton

International Centre for Spatial Computational Learning
EPSRC EP/S030069/1

Motivation



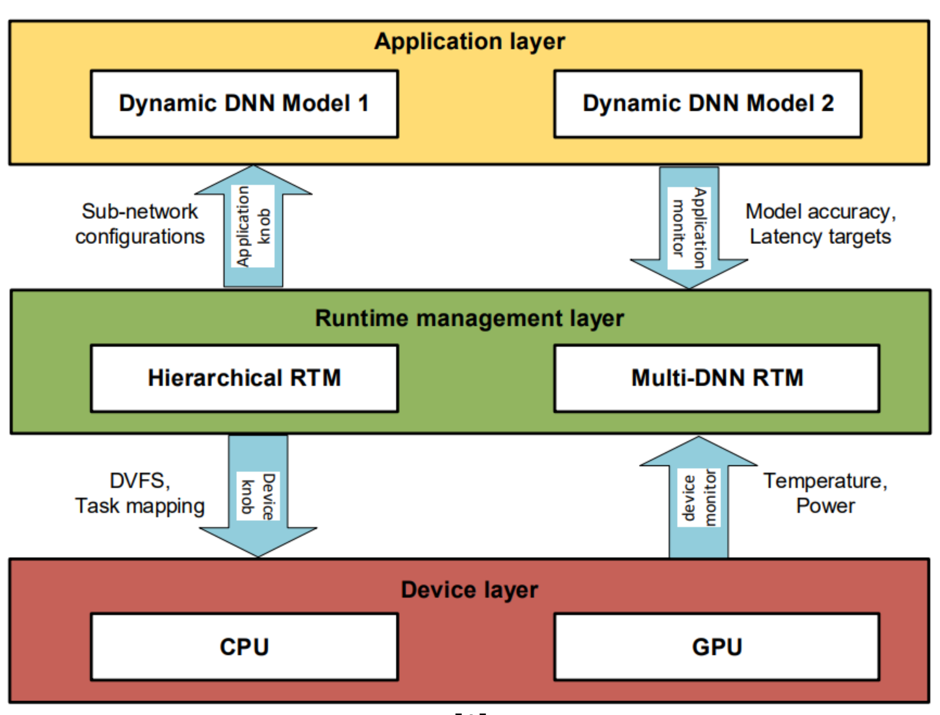
[1]



[2]

- DNN inference is increasingly being executed on mobile and embedded devices, thanks to its low latency and improved privacy. However, DNN models are both computationally and memory access intensive.
- Efficient deployment of DNN models faces three primary challenges:
  1. **[Hardware Variability]** Achieving consistent performance across platforms is difficult due to significant variations in hardware computing capabilities.
  2. **[Application Variability]** A single DNN model (e.g., machine translation) can be utilized in various applications (e.g., real-time speech translation, text translation), but their performance requirements differ.
  3. **[Runtime Variability]** The hardware resources available to the model may change during runtime due to factors such as thermal throttling or the DNN model sharing resources with other applications.

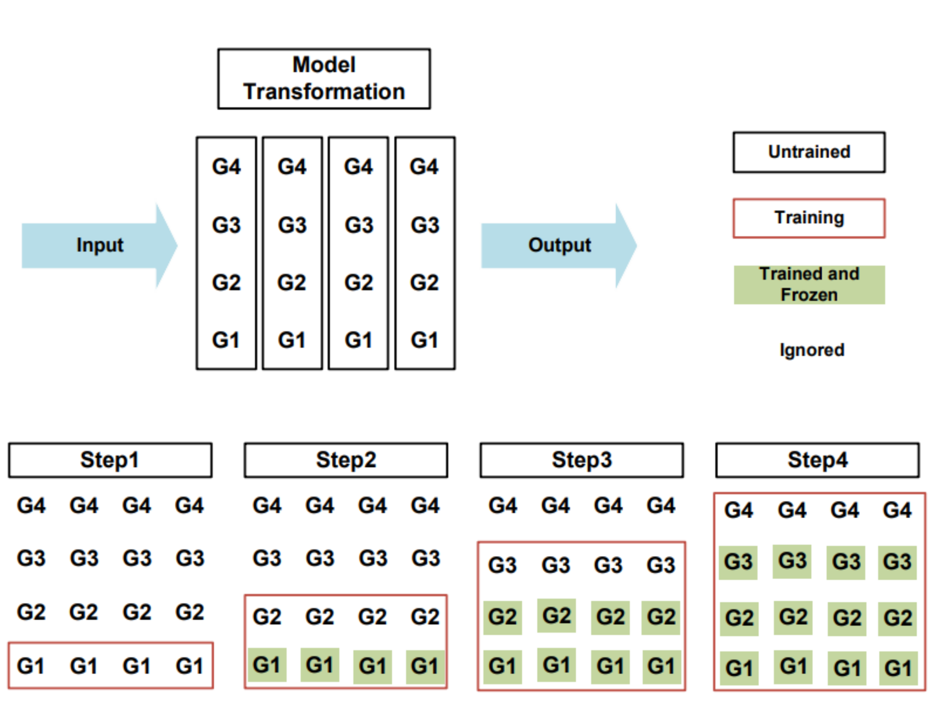
Runtime Algorithm and Hardware Management [1,5]



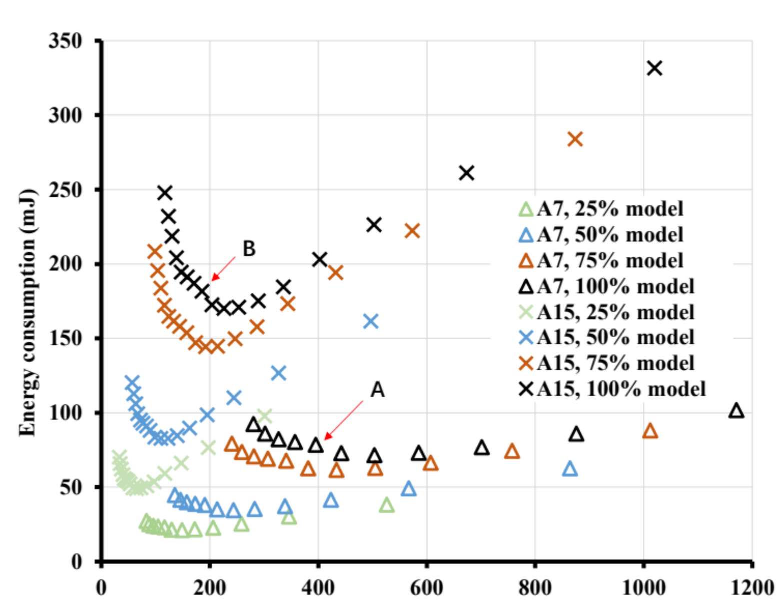
[1]

- Traditional runtime resource management primarily focuses on hardware adjustments (e.g., DVFS, task mapping), treating DNN models as general applications and overlooking domain-specific optimization opportunities.
- In our research, we have developed dynamic neural networks that facilitate runtime adjustments for both algorithms and hardware.

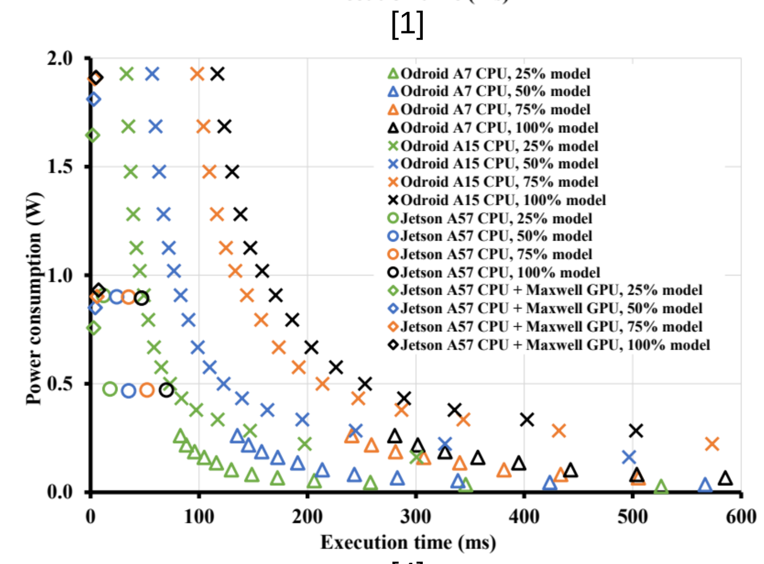
Incremental Training and Group Convolution Pruning [1,3,4]



[3]



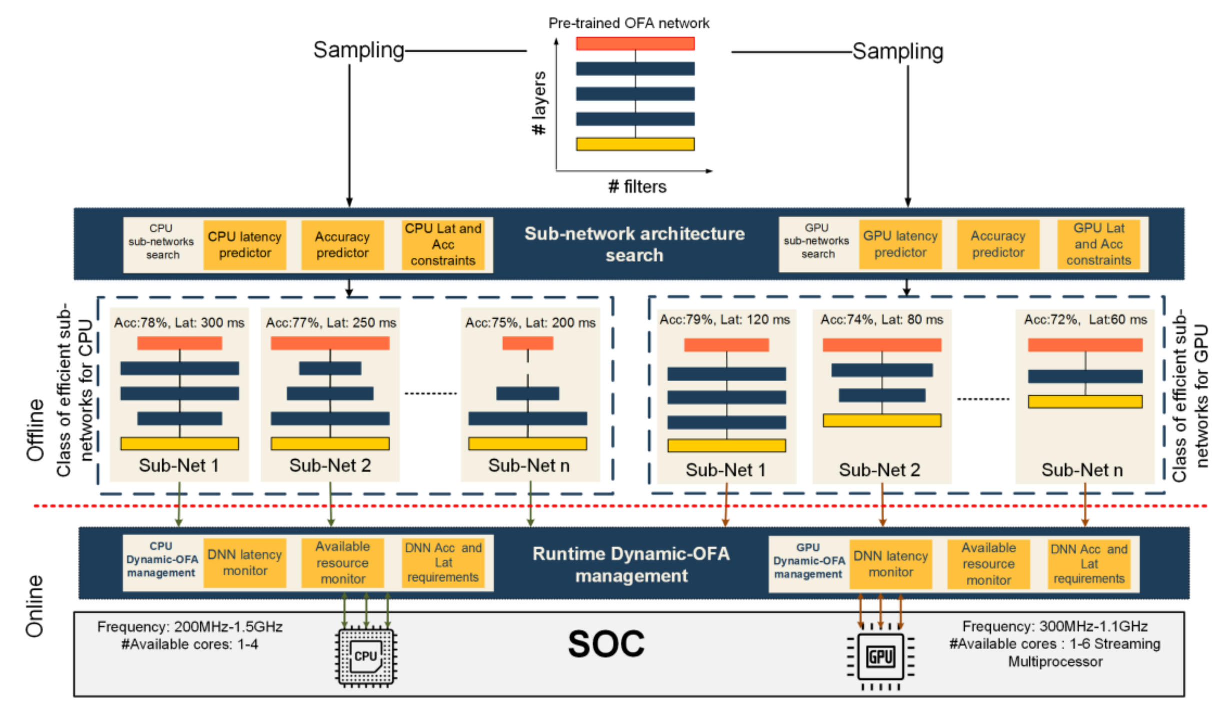
[1]



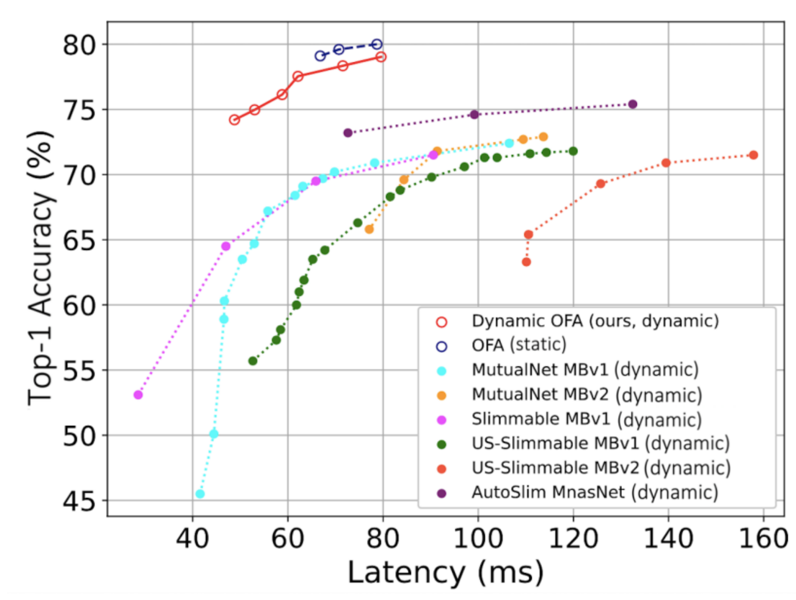
[4]

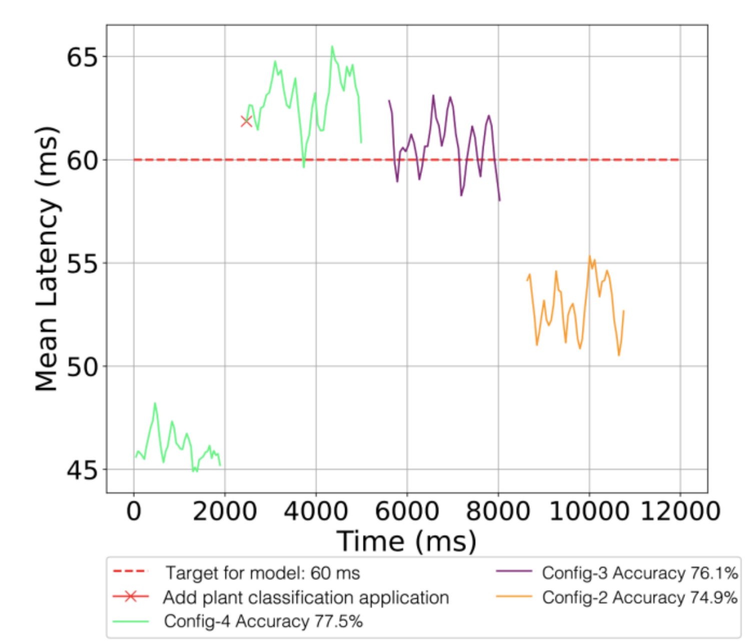
- Convolution layers are divided into groups, which are trained incrementally.
- A dynamic neural network with four sub-network configurations is created. Each sub-network offered unique accuracy, latency, and power/energy trade-offs.

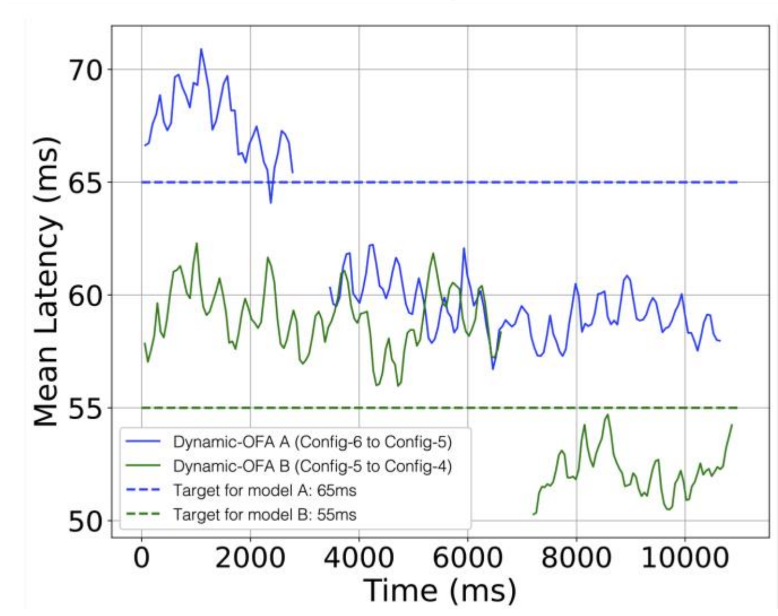
Dynamic Supernetwork [2, 6]

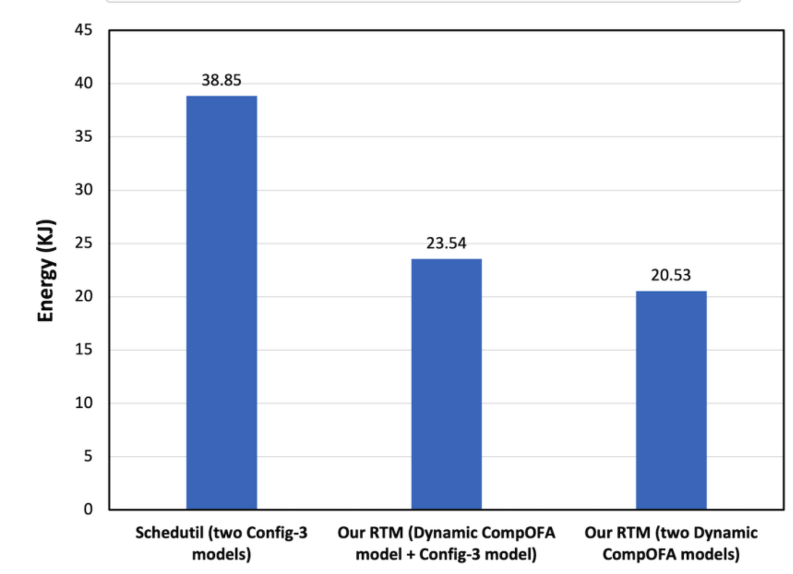


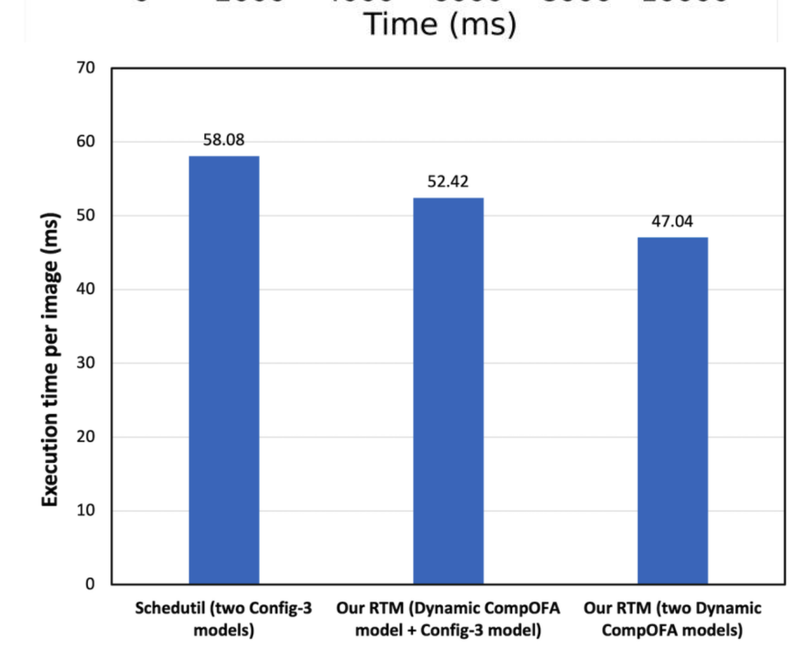
- Sub-networks are sampled from a pre-trained supernetwork, and a dynamic neural network is created using the sub-networks located on the Pareto-front of performance trade-off.
- The sampling process is conducted separately for CPUs and GPUs, as the most efficient sub-network architectures differed for these heterogeneous computing resources. Each sub-network provided unique accuracy, latency, and power/energy trade-offs.
- During runtime, the sampled sub-networks can be switched to meet the desired performance targets, adapting to the time-varying available hardware resources.

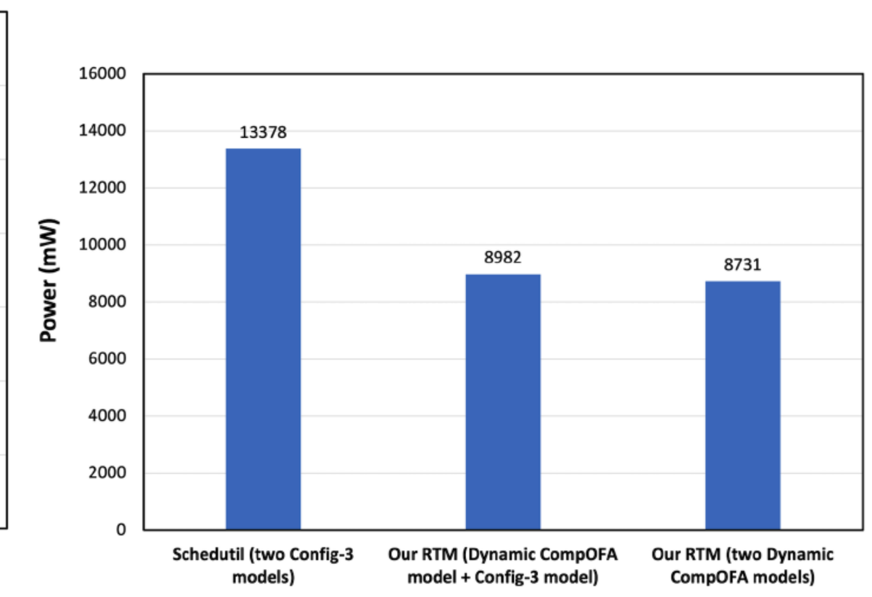












Conclusion

- Our research addresses the challenges associated with the efficient deployment of DNN models on heterogeneous computing platforms.
- We have developed novel dynamic neural network methods for both static models and supernetworks.
- The proposed system framework offers great performance trade-off adaptability, and system efficiency through runtime resource management, which facilitates runtime adjustments for both algorithms and hardware, enabling system adaptation to the dynamic performance targets and available hardware resources.

Reference

[1] Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett. Optimising Resource Management for Embedded Machine Learning. In Design, Automation and Test in Europe Conference (DATE), 2020.
[2] Wei Lou*, Lei Xun*, Amin Sabet, Jia Bi, Jonathon Hare, and Geoff V Merrett. Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms. In Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021.
[3] Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett. Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms. In ACM/IEEE 1st Workshop on Machine Learning for CAD (MLCAD) , 2019.
[4] Lei Xun, Bashir M Al-Hashimi, Johnathan Hare, and Geoff V Merrett. Runtime DNN Performance Scaling through Resource Management on Heterogeneous Embedded Platforms. In tinyML EMEA Technical Forum, 2021.
[5] Lei Xun, Bashir M Al-Hashimi, Jonathon Hare, and Geoff V Merrett. Dynamic DNNs Meet Runtime Resource Management on Mobile and Embedded Platforms. In 4th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK), 2022.
[6] Hishan Parry, Lei Xun, Amin Sabet, Jia Bi, Jonathon Hare, and Geoff V Merrett. Dynamic Transformer for Efficient Machine Translation on Embedded Devices. In ACM/IEEE 3rd Workshop on Machine Learning for CAD (MLCAD) , 2021.