tinyML EMEA Technical Forum 2021
June 7-10, 2021

Runtime DNN Performance Scaling through Resource Management on Heterogeneous Embedded Platforms

Lei Xun, PhD Candidate, University of Southampton, UK

Advisor: Geoff V. Merrett, Jonathon Hare, Bashir M Al-Hashimi

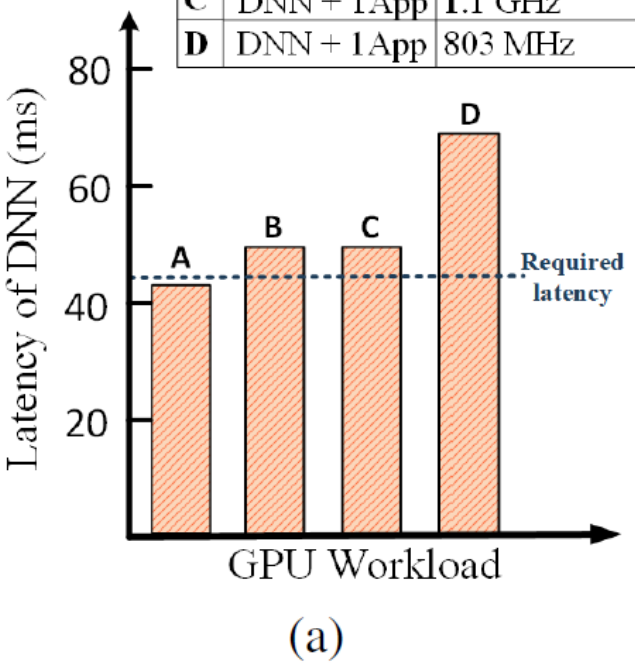
International Centre for Spatial Computational Learning (EPSRC EP/S030069/1)

June 10, 2021

Motivation for dynamic DNNs

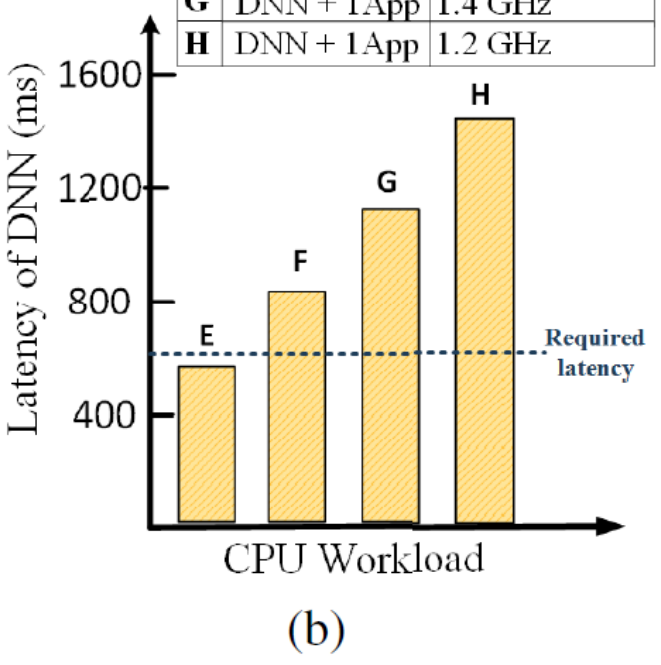
- DNNs are typically compressed before deployed on embedded platform
- However, the assumed hardware resources may not be available at runtime

	Workload	GPU freq
A	DNN	1.1 GHz
B	DNN	803 MHz
C	DNN + 1App	1.1 GHz
D	DNN + 1App	803 MHz



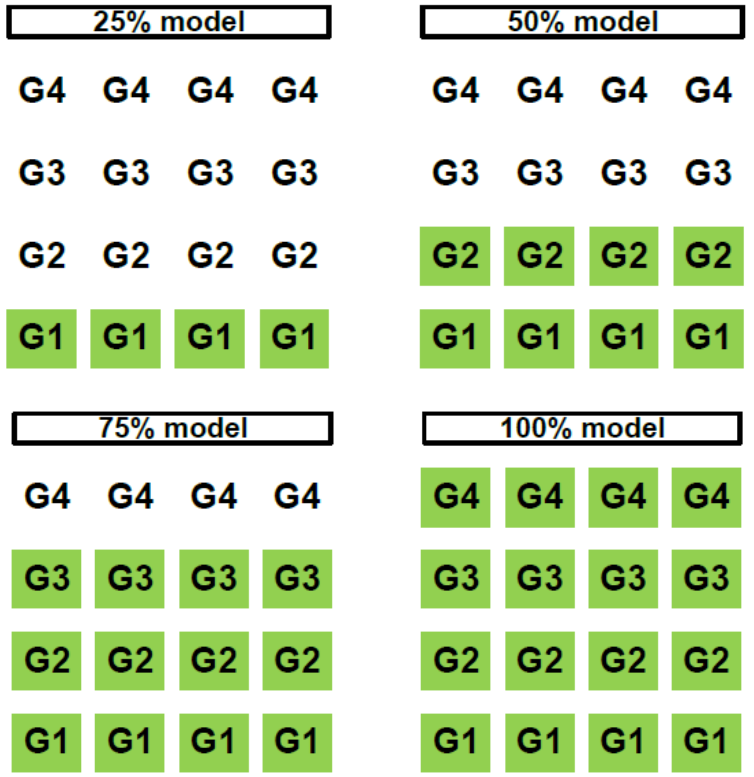
(a)

	Workload	CPU freq
E	DNN	1.4 GHz
F	DNN	1.2 GHz
G	DNN + 1App	1.4 GHz
H	DNN + 1App	1.2 GHz



(b)

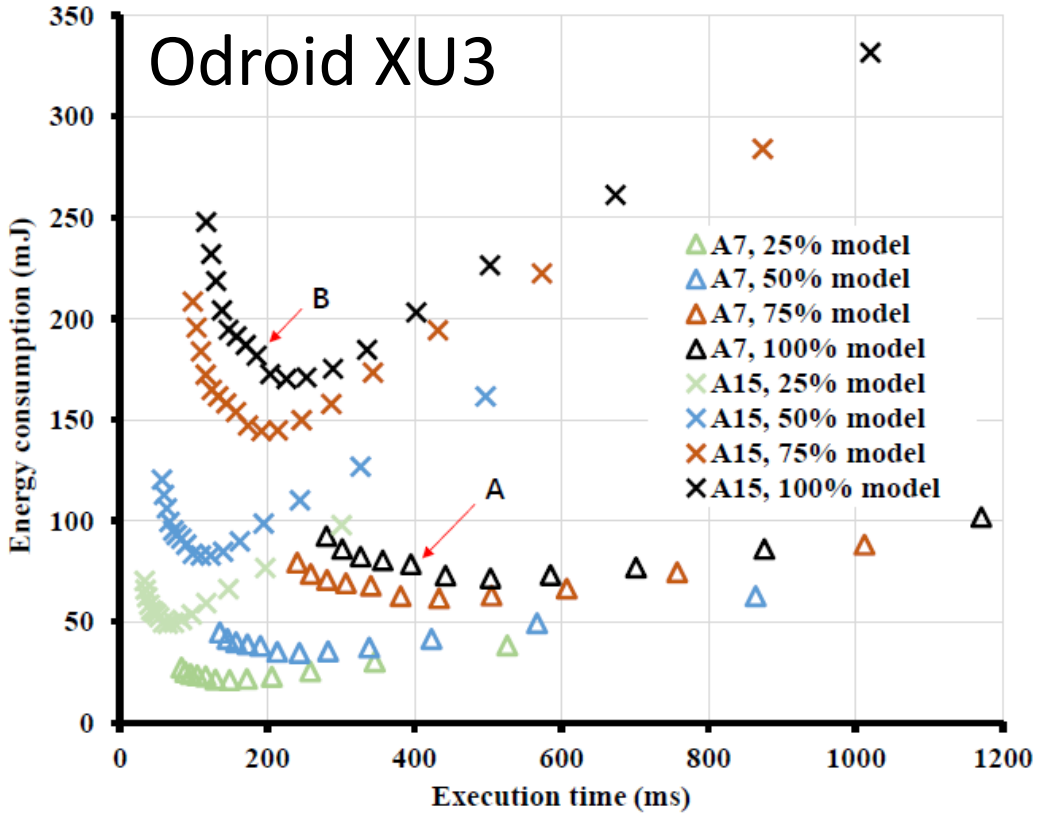
- Dynamic DNNs can be executed partially to trade-off accuracy for latency/power/energy reduction

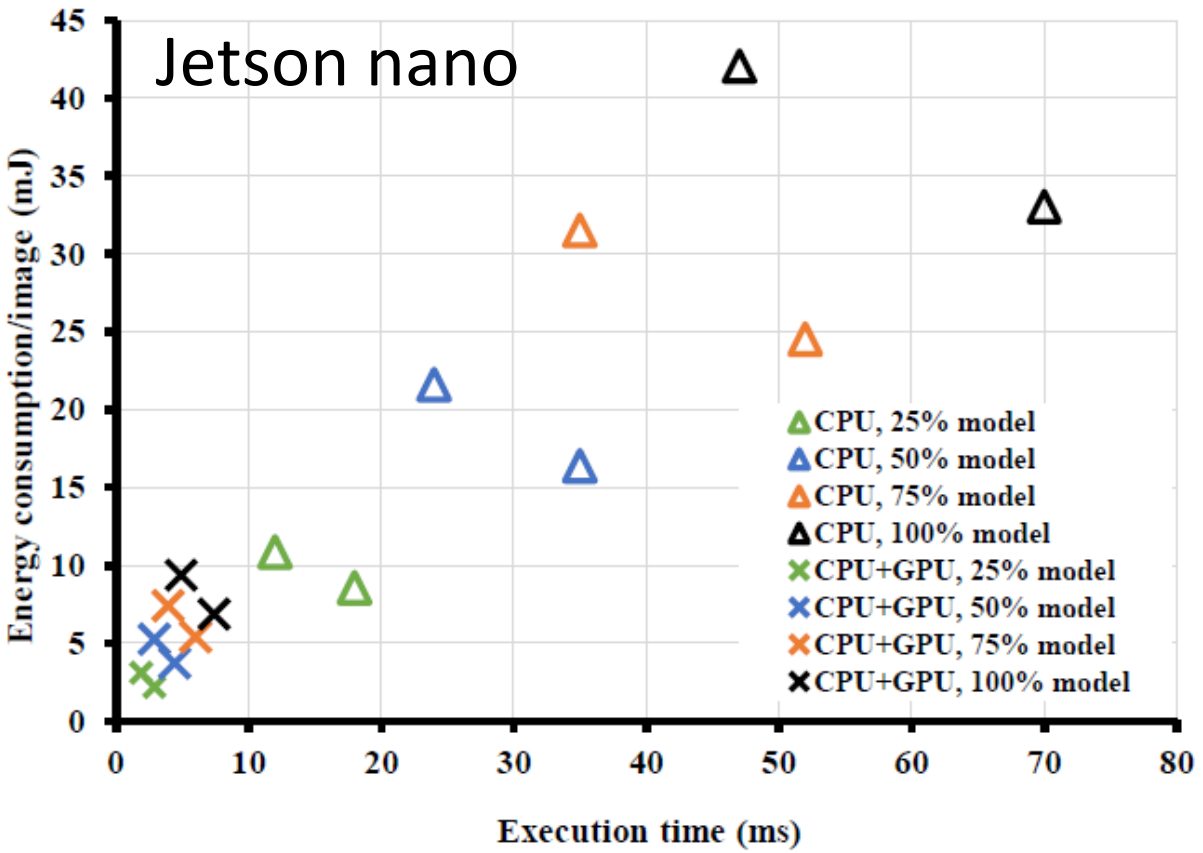


L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, "Incremental training and group convolution pruning for runtime DNN performance scaling on heterogeneous embedded platforms," in *Workshop on Machine Learning for CAD (MLCAD)*, 2019.

Accuracy/latency/energy trade-offs

- Subnetworks are shown in different colors, computing elements are shown in different symbols and frequency scaling are shown in points
- Operating points example: On Odroid XU3, A has the best trade-off under 100mJ and 400ms requirements, B is the best for 200mJ and 200ms





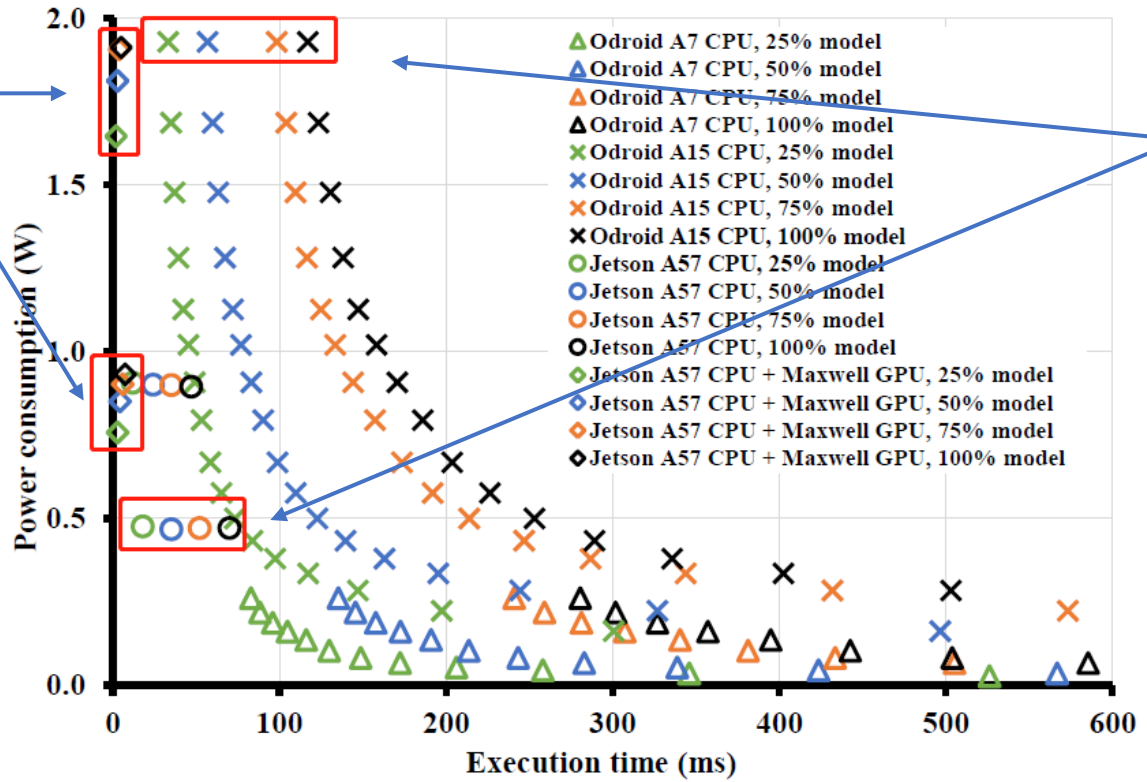
L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, "Optimising Resource Management for Embedded Machine Learning," in *Design, Automation & Test in Europe Conference (DATE)*, 2020.

Accuracy/latency/power trade-offs

Now we know that dynamic DNNs can trade-off accuracy for latency and energy, what about power?

The power of GPU

- Scales with frequency scaling
- Scales with dynamic DNN, this provide us new opportunities to meet power target



The power of a single-core CPU:

- Scales with frequency scaling
- Does not scale with dynamic DNN, since the computation intensity does not change when using smaller subnetworks, only the latency changes due to less amount of works

To know more about dynamic DNNs, please check out our SOTA dynamic DNN paper:
W. Lou, L. Xun, A. Sabet, J. Bi, J. Hare, and G. 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. https://arxiv.org/abs/2105.03596

Thank you

Lei Xun https://www.linkedin.com/in/lx2u16/

International Centre for Spatial Computational Learning https://spatialml.net/

Reference:

[1] L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, “Incremental training and group convolution pruning for runtime DNN performance scaling on heterogeneous embedded platforms,” in *Workshop on Machine Learning for CAD (MLCAD)*, 2019.

[2] L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, “Optimising Resource Management for Embedded Machine Learning,” in *Design, Automation & Test in Europe Conference (DATE)*, 2020.

[3] W. Lou, L. Xun, A. Sabet, J. Bi, J. Hare, and G. 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.