Characterizing DNN Models for Edge-Cloud Computing

Chunwei Xia[†§], Jiacheng Zhao [†], Huimin Cui [†§], Xiaobing Feng [†§]

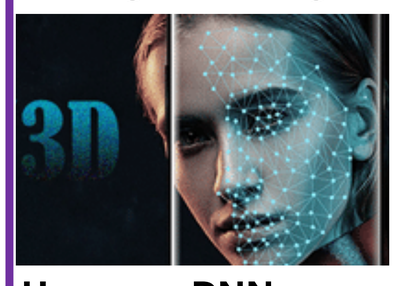
[†]*State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences*

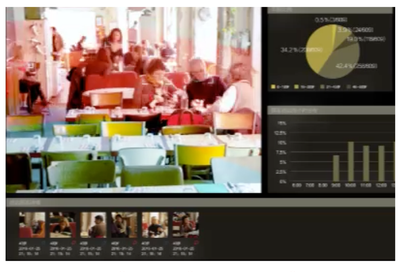
[§]*School of Computer and Control Engineering, University of Chinese Academy of Sciences*

Beijing, China

Introduction

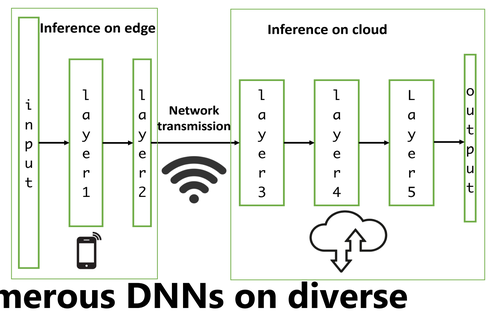
Deep Neural Networks (DNNs), deployed in the cloud nowadays, are moving to the edge, e.g. mobile phones.





However, DNNs are computation-intensive and edge devices are always resource-constrained, driving us to run the DNNs between the edge and the cloud collaboratively, i.e. Edge-Cloud Computing.

Deployment	Device Only	Cloud Only	Edge-Cloud
Computation Latency	✗	✓	✓
Network Transmission	✓	✗	✓



But, the performance characteristics of numerous DNNs on diverse platforms are not clear, especially on edge devices.

DNN models	Perf. Characteristics	Diverse Platforms
Multi-Add Count	Computation Latency	CPU cores/freq
Model Size	Memory Occupation	Memory
DNN Operations	Energy Consumption	Battery Life

Goals:

To provide insights on performance characteristics of representative DNNs on cloud and edge platforms

Detailed Characteristics

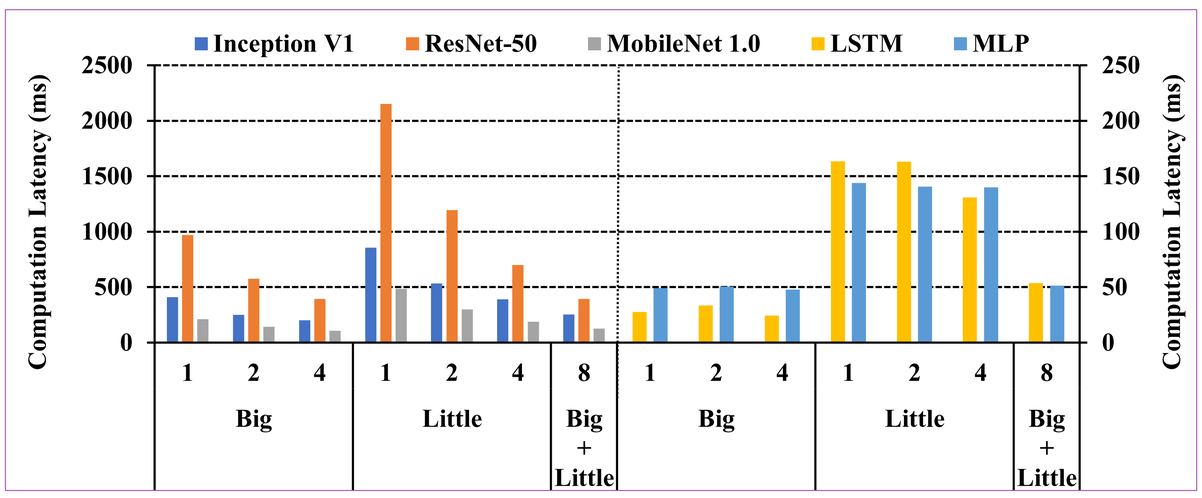


Fig. 2: Computation latency of five DNNs on Edge CPU A with 1, 2, 4, 8 threads when using Big and/or Little cores.

More Threads = Lower Latency ? (CNNs: Yes, LSTM/MLP: Not Exact)

More Resources != Lower Latency for big.LITTLE architecture.

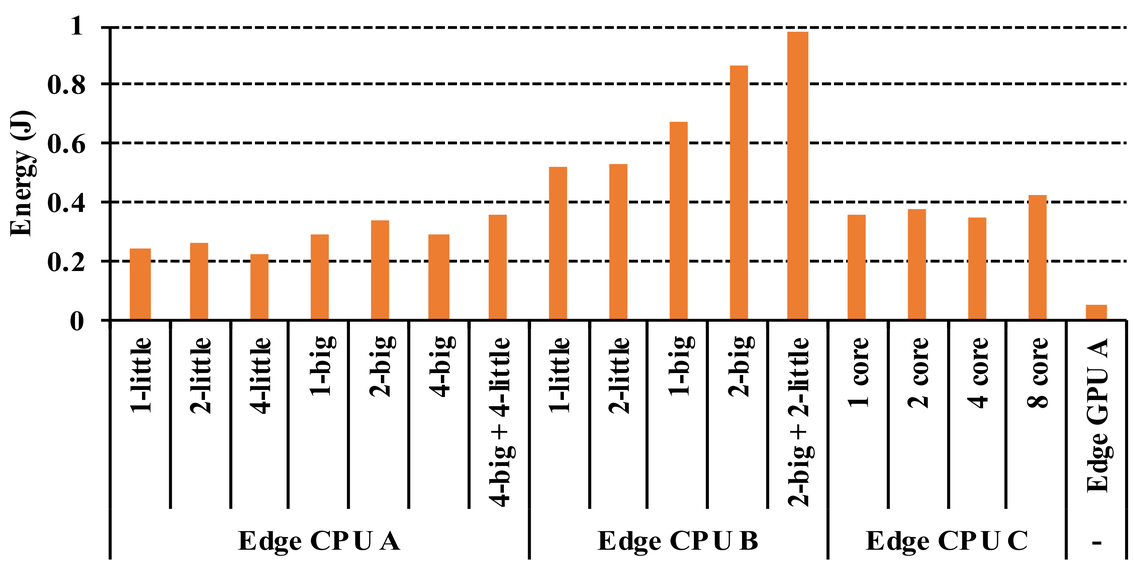


Fig. 3: Comparison of edge CPU and GPU energy consumption for MobileNet 1.0

Balance Latency and Energy consumption via thread-to-core mapping.

Experimental Setup

DNN Models:

Category	Name	Input	Output	Layers	# Params	FLOPs
CNNs	Inception V1	[224,224,3]	[1000]	22	6.79M	3.19B
	ResNet-50			50	25.6M	3.8B
	MobileNet 1.0			15	4.2M	576M
RNNs	LSTM	[20, 200]	[20, 10000]	2	2.65M	14.8M
MLP	MLP	[1,784]	[10]	5	13.9M	13.9M

Cloud & Edge Platforms:

Cloud		Edge		
Platform	Notation	Platform	SoC	Notation.
E5-2620 v4	*Server CPU A*	OnePlus 5T	Sd. 835	*Edge CPU A*
E5-1603 v4	*Server CPU B*	OnePlus 3	Sd. 820	*Edge CPU B*
Tesla K40c	*Server GPU A*	Redmi Note 4x	Sd. 625	*Edge CPU C*
GTX 1070	*Server GPU B*	Jetson TX2	Pascal GPU	*Edge GPU A*

benchmark_model of Tensorflow for performance analysis

Overall Evaluations

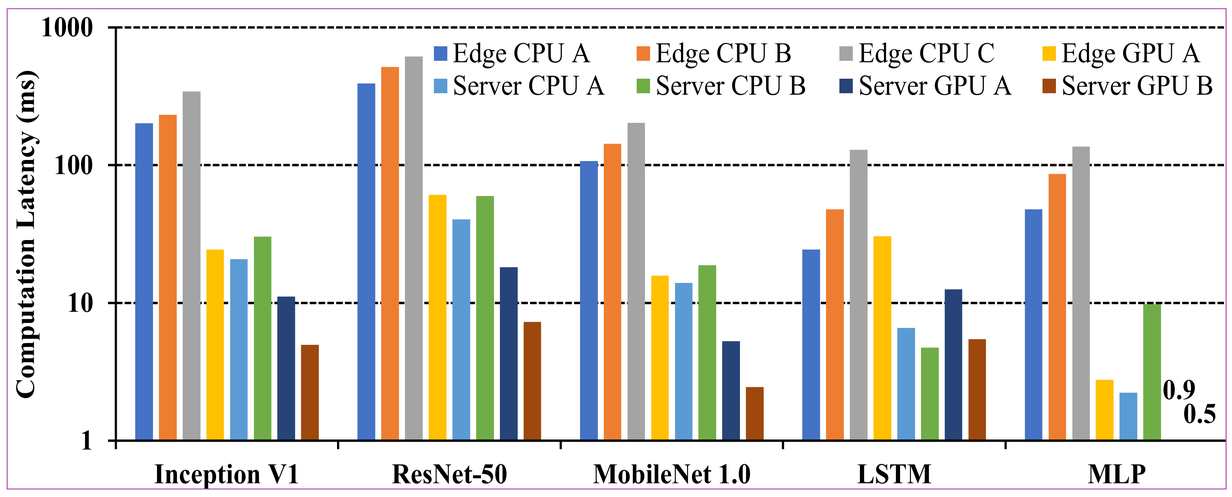


Fig. 1: Computation latency of five DNN models on eight platforms with optimal configuration

Edge CPU: Sufficient for LSTM/MLP/MobileNet, not Inception/Resnet.

GPU (Server or Edge) VS. CPU: 5x faster for CNNs and MLP, 1.0x for LSTM

Conclusion

Diversity of platforms leads to much variance in computation latency

Dynamic deployment strategy shall be designed for big.LITTLE arch.

Much efforts are required to balance latency and energy consumption of DNNs for edge devices.

Future work: More edge accelerators, e.g. DSP, NPU; Automatic computation partition.

xiachunwei@ict.ac.cn

Get this e-copy