DeepEdge – Deep Neural Networks on Edge Devices

Srujana Malisetti, Akanksha Raina, Rohit Singh
Computer Science Department
The University of Georgia
{Srujana.Malisetti, Akanksha.Raina, Rohit.Singh}@uga.edu

Introduction

Ever wondered how Facebook auto-tag feature works or Alexa interacts with you? Technology behind these slackening features is Deep learning. Most of these Intelligent assistants' state-of-the-art processes pass on computations to the cloud and get the results for users, as in below figure.
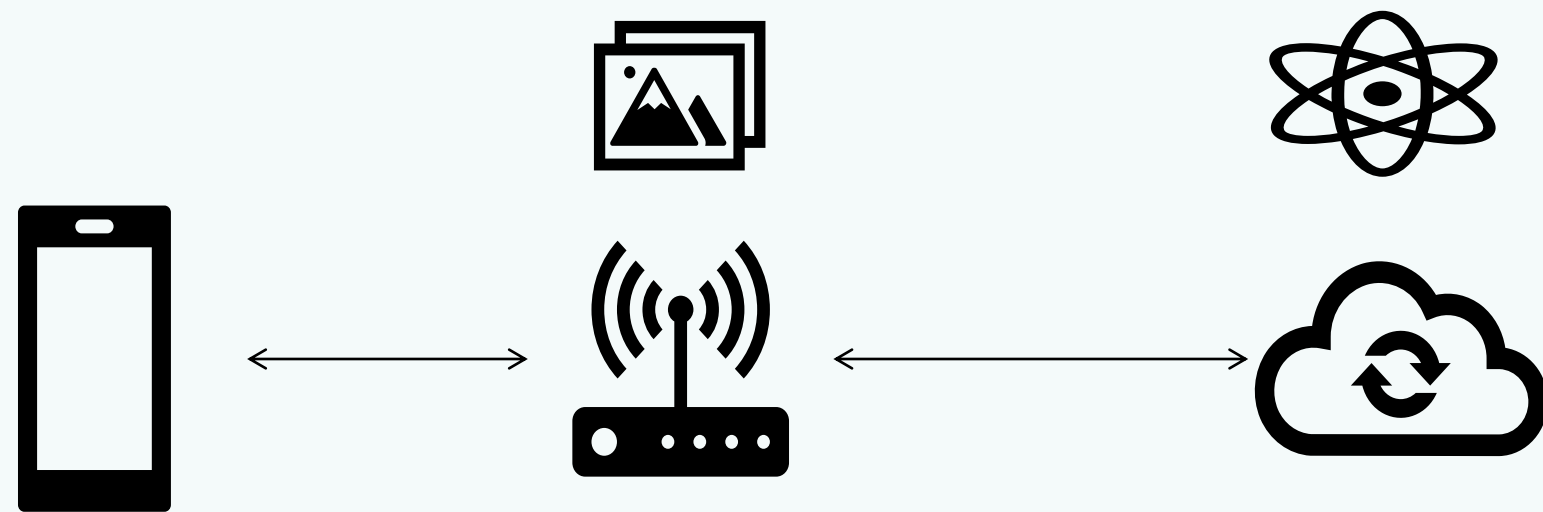
Problems to be noted in this approach is data transfer overhead in terms of latency and energy costs when transferred over the Internet . As we say Intelligent assistants with SoC integration are sharpening up to perform the computations without passing on to the cloud. Our research focuses on leveraging this power on various Edge devices and run computations on devices.
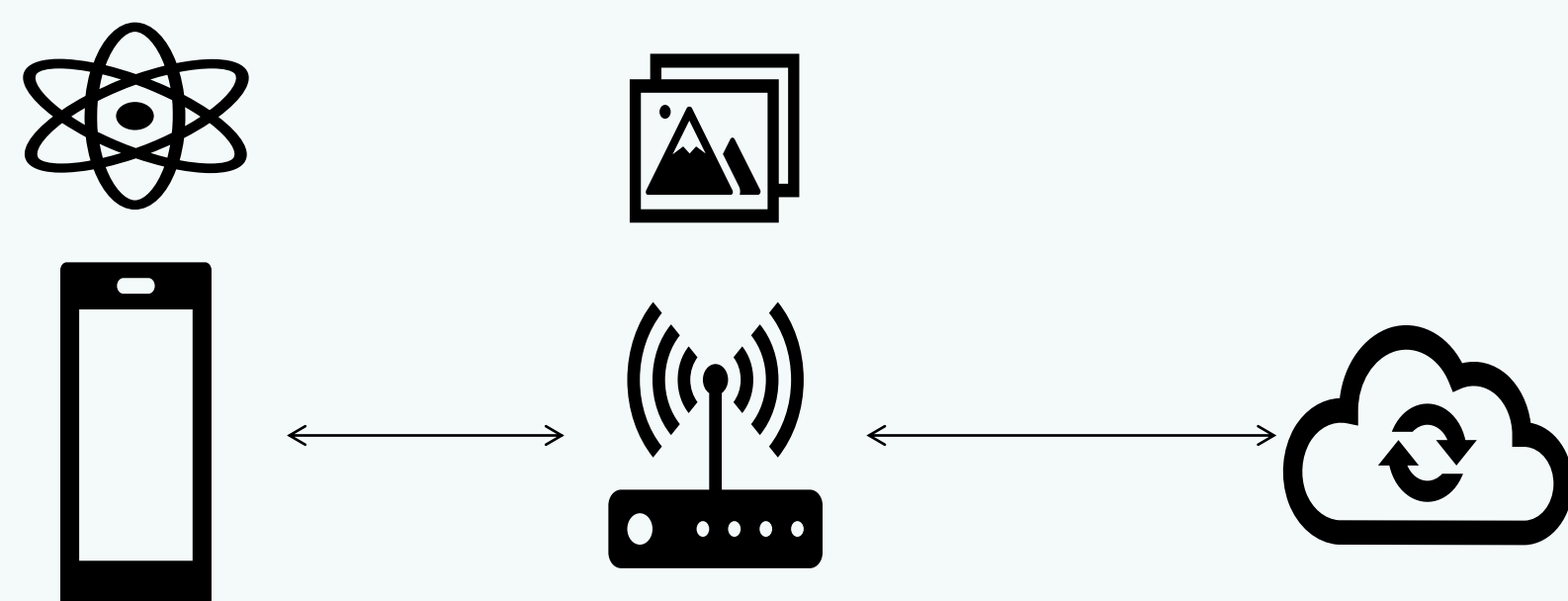
Related Work

- Neurosurgeon [1], deals investigating on finding the optimized partition point to divide the computations between edge devices and cloud servers. Their results are as below,

Across 8 benchmarks		Average	Maximum
Latency		3.1x	40.7x
Mobile energy Consumption		59.5%	94.7%
Datacenter Throughput		1.5x	6.7x

- Distributed Perception by Collaborative Robots [2], deals investigating on distributing the computations among group of Raspberry Pi edge devices.

Approach

| Deploy ML Frameworks Caffe, TensorFlow Lite |
|---|

Deploy Deep Neural Networks AlexNet, GoogleNet

Evaluate Latency, Memory Consumption, Energy Consumption

Devices chosen *Android Mobile Phone, Laptop with NVIDIA Drivers, Raspberry Pi, Jetson Nano.*

Contributions

- We show which device performs efficiently considering Latency, memory and energy consumptions for different DNNs.
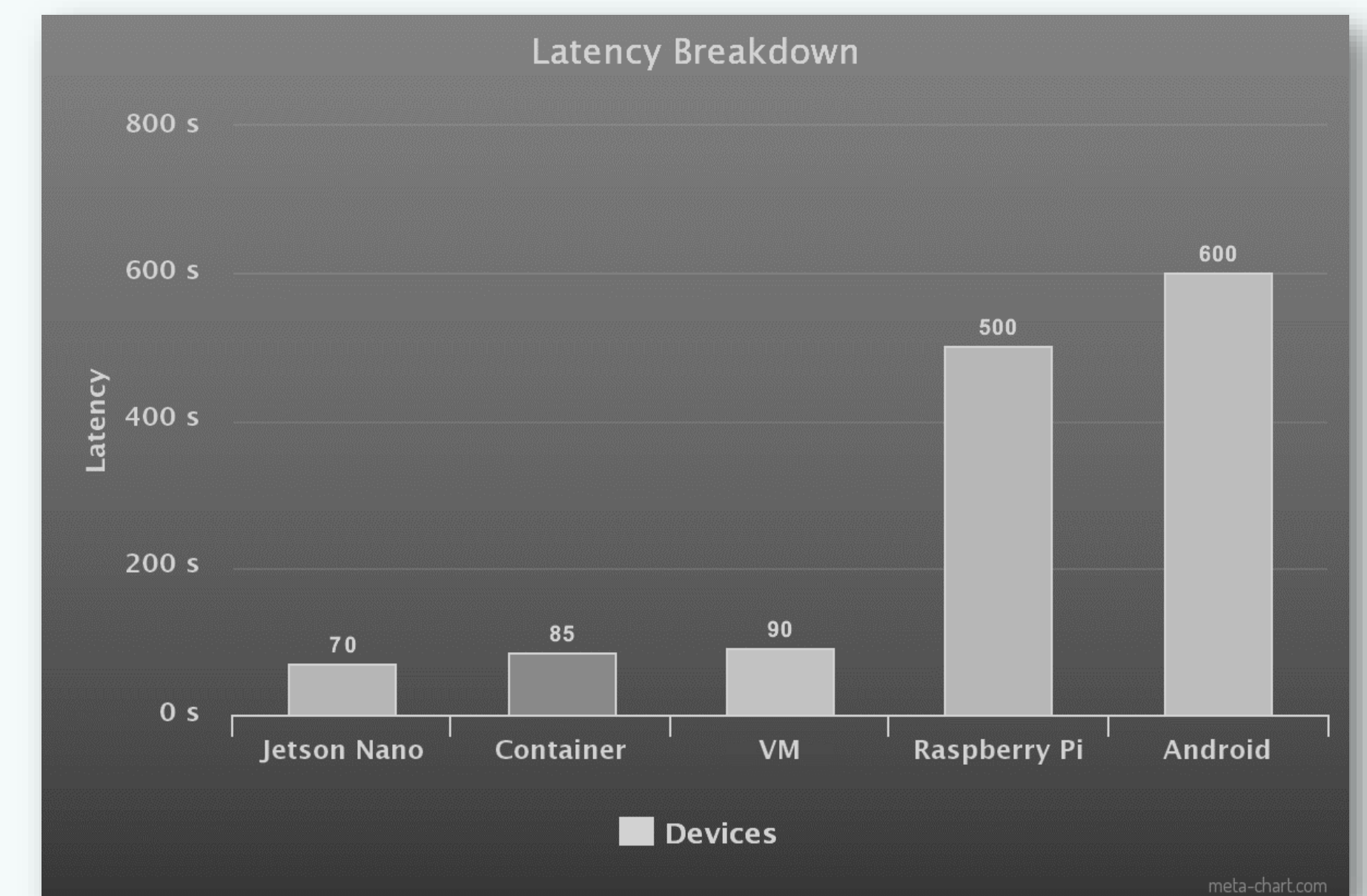
References

1. Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," in 2017 ASPLOS, 2017.
2. Ramyad Hadidi , Jiashen Cao , Matthew Woodward, Michael S. Ryoo, and Hyesoon Kim, "Distributed Perception by Collaborative Robots" in 2018, 2018 IEEE.

Acknowledgments