Tootiya Giyahchi

Aaron Bae

Ritwik Nandakumar

**CompSci 237 - Spring 2020: Distributed Systems Middleware**

**Project Proposal**

Nowadays Deep Neural Network methods are widely adopted in IoT applications. Due to the high compute-intensive nature of DNNs, the workload of such applications must be offloaded to the cloud or edge. Using edge nodes are especially beneficial for vision and object recognition applications (e.g. google Glasses and DeepLens, AR applications, robotics, unmanned aerial vehicle (UAV), and E-healthcare system) where a great amount of data should be transmitted and cloud's communication delay would reduce efficiency. Edge computing provides efficient data processing by minimizing the amount of long-distance communication between IoT devices and the cloud server. Therefore edge nodes can reduce bandwidth occupation, latency, and energy consumption[1].

Hence, edge devices have become a hub for real-time DNN based applications and many application-specialized DNN accelerators have been developed on the edge.

[2], [3] and [4] also propose how Edge nodes can help with running heavy DNNs for different applications such as vision[4], IPA [3] and object detection [2], so improving and designing edge nodes for such purposes has become very important.

Since the edge is a shared commodity, there might be different applications with different requirements among end devices that may need acceleration on the edge.

In this scenario, there will be a queue of tasks demanding real-time responsiveness from the edge and the edge has to meet each application's requirements and yet it is not unlimited in computing resources.

So it seems necessary to come up with some strategies for edge devices to prioritize tasks and decide to allocate the best acceleration options that meet the requirements and maximize the utilization without adding overheads.

On the other hand, for IoT vision applications we expect tasks to arrive with regularity as the images are being sent from devices at certain rates (but due to network uncertainties, it is not a fixed pattern). An edge middleware can exploit such regularity to optimize resource allocation and manage tasks efficiently.

In this project, we would like to examine strategies for extracting a pattern from the queue of incoming requests to the edge as a part of middleware between applications and edge devices.

**Agenda:**

* A quick survey on state-of-the-art IoT middleware tools: AWS Greengrass, Microsoft Azure IoT
* Define the problem from IoT and network perspective
* Pick tools and methods to implement our setup
* Experiment the opportunity of pattern extraction from the incoming requests to the edge

References:

[1] W. Yu *et al*., "A Survey on the Edge Computing for the Internet of Things," in *IEEE Access*, vol. 6, pp. 6900-6919, 2018.

[2] Jridi, M.; Chapel, T.; Dorez, V.; Le Bougeant, G.; Le Botlan, A. SoC-Based Edge Computing Gateway in the Context of the Internet of Multimedia Things: Experimental Platform. J. Low Power Electron. Appl. 2018, 8, 1.

[3] Amir Erfan Eshratifar and Massoud Pedram. 2018. Energy and Performance Efficient Computation Offloading for Deep Neural Networks in a Mobile Cloud Computing Environment. In Proceedings of the 2018 on Great Lakes Symposium on VLSI (GLSVLSI '18).

[4] ao, Cong et al. “FPGA/DNN Co-Design.” Proceedings of the 56th Annual Design Automation Conference 2019 on - DAC ’19 (2019).