COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 8

Re-thinking computation offload for efficient inference on IoT devices with duty-cycled radios – Summary

The problem that is addressed in this paper is about the efficient execution of deep neural networks on low power IoT platforms, mainly the ones with duty-cycled radios. This is addressed as an important obstacle because IoT devices have limited computational resources and they need to prolong their battery life by conserving energy. Efficient execution of these DNNs allow tasks like object detection in IoT settings, facilitating applications such as security monitoring, environmental sensing and smart home automation. The hypothesis of the paper is that through offloading early exit computation to the cloud, it enables to achieve lots of performance benefits in terms of both energy consumption and execution latency for the IoT cloud systems.

The authors propose a solution for this, which is called Fast and Light Emerging Execution on Tiny devices also known as FLEET. This solution mainly helps with the utilization of cloud resources to assist in early-exit computation, which reduces computation cost of the model on resources constrained IoT devices. FLEET minimizes energy consumption and latency without compromising on the accuracy. This is done by stopping the inference process early within the firs few layers when possible. One major limitation of this proposal is that there might be dependency on network connectivity when it comes to cloud assistance, If the network connectivity is unreliable, then the performance of FLEET maybe impacted. In order to have a work around this, integrating mechanisms for local cache of cloud assisted models could help by ensuring that the system remains functional even if there is no stable network connection.

Srimathi Mahalingam