Name: **Mateen Ulhaq**
SFU faculty/major: Engineering Science; Engineering Physics Honors

Title of presentation: Shared Mobile-Cloud Deep Learning Model Inference

Abstract

As AI applications for mobile devices become more prevalent, there is an increasing need for faster inference on mobile. Inference is the process of taking input data (e.g. images, audio), processing it through a deep learning model, and retrieving a result (e.g. a description of the image/audio contents).

Currently, models are either run completely on the mobile device or completely on the cloud. However, running inference only on the cloud costs network bandwidth introduces latency, and requires the input data to be fully transferred to the cloud, creating privacy concerns.

We demonstrate an alternative approach: shared mobile-cloud inference. Partial inference is performed on the mobile device to reduce the size of the input data. The results of partial inference are then transmitted to the server for further inference. This strategy can improve latency, reduce bandwidth usage, and provide privacy protection because the input data never leaves the mobile.