Adaptive Routing of Requests to Distributed Analytics Models

Shiqiang Wang (IBM US), **Dave Conway-Jones** (IBM UK), **Stephen Pasteris** (UCL), **Kevin Chan** (ARL), **Mark Herbster** (UCL)

Introduction

In tactical coalition environments, data are collected by sensors distributed in the tactical field. Such data can be used either directly or indirectly by analytics applications.

One common approach in analytics applications is to train a model using the locally collected data, so that the model can be further used for specific tasks, such as image classification.

Challenges

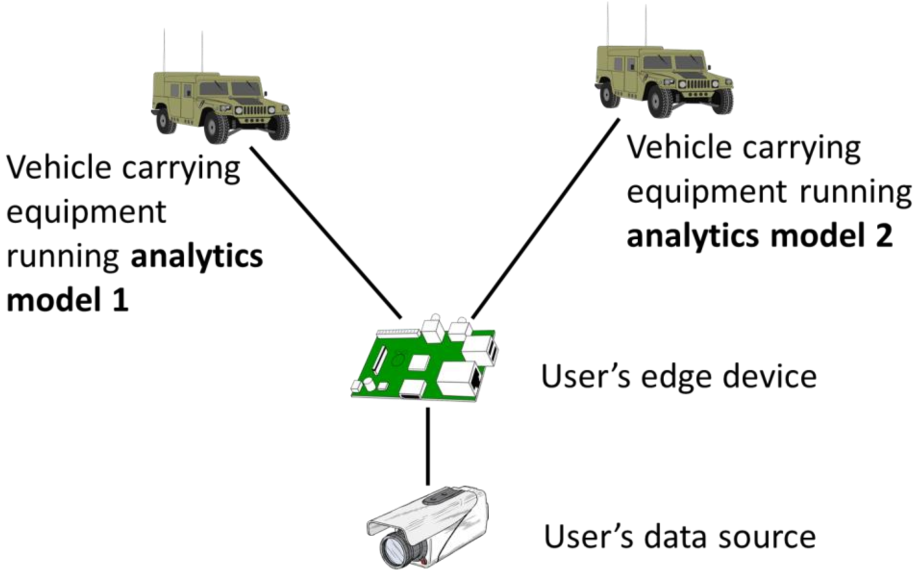
- Due to the locality of datasets, models trained on different nodes usually have different capabilities
  - For example, different models may be good at differentiating different categories of images
- The sharing of local models or dataset can be infeasible in a tactical coalition system
  - Limited communication bandwidth
  - Security considerations
    - Both the model and the dataset can pose risks of leaking sensitive information

Proposed System

- Distributed analytics is performed in an on-demand manner
- Models and datasets remain at local nodes
- Analytics requests (such as images that need to be classified by the model) are generated by users in the tactical field
- When a request is generated by the user, the request is routed to a suitable node to process it
- After processing, the result is sent back to the user

Demo Setup

- Two computational nodes that run different analytics models on them
- A user operating an edge device has different images and would like to detect objects in those images
  - The user does not know which analytics model suits the best for its task
- A cost is incurred when user queries a model
  - The cost is related to the communication and computation resource consumption
  - Querying different models generally incurs different costs, because the models are located on different computational nodes
- The system decides which model to route the user's input image to
  - Based on a preliminary analysis of the input image on the edge device
  - The image may be routed to multiple models for analysis before the final result is obtained
- The policy specifying the routing of images to models can be *updated in real time and in an online manner*
  - Based on historical observations of user data input and the behavior of different models


Vehicle carrying equipment running **analytics model 1**
Vehicle carrying equipment running **analytics model 2**
User's edge device
User's data source

Distributed Analytics and Information Science
International Technology Alliance
Annual Fall Meeting, September 2018