Adaptive Routing of Requests to Distributed Analytics Models

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

Scope: Models trained on datasets at local nodes process analytics requests that can originate from remote nodes. The system adaptively routes requests to nodes with suitable models for processing.

Description:

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 is to train a model using the locally collected data, so that the model can be further used by analytics applications, such as image classification. 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, due to the limited communication bandwidth as well as security considerations, as both the model and the dataset can pose risks of leaking sensitive information.

In this work, we consider the scenario where distributed analytics is performed in an on-demand manner. The models and datasets remain at the local nodes. Analytics requests are generated by users in the tactical field. When a request is generated, the request is routed to a suitable node to process it.

We demonstrate an emulated system with 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, but it does not know which analytics model suits the best for its task. A cost is incurred when the user queries a model, where 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.