P4 Demo: Unsupervised detection of anomalous model inputs at and emergence of new classes at the edge

Contributors: Nirmit Desai, Raghu Ganti (IBM-US)
Heesung Kwon (ARL)
Ian Taylor (Cardiff)

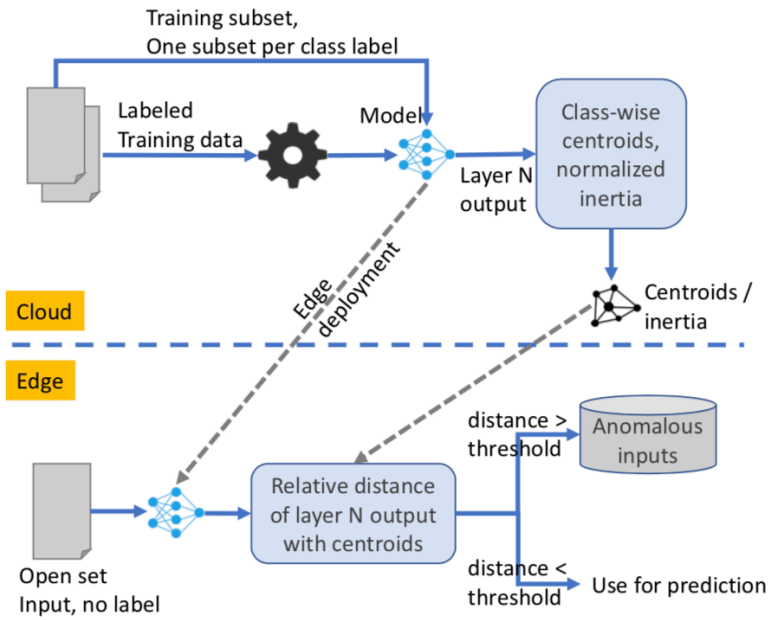
Scope: Demonstrating how unlabelled model inputs at the edge can be classified as anomalous when they not match the model domain and how new classes can be discovered from the anomalous inputs

Equipment: Power supply, display screen

Description:

With growing applications of machine learning across many domains, it is increasingly important to assess whether or not a given machine learning model is applicable to a domain. It is well-known that accuracy of a deployed model is highly dependent on the "similarity" of the deployed environment to the training environment where data was collected and processed. Present methods for assessing the applicability of a given model on a new dataset rely on the dataset having the "ground truth" labels. However, in many scenarios, especially those involving coalition edge operations, labeling the new datasets is costly or not feasible. As a result, model predictions based on anomalous data inputs may lead to erroneous predictions and faulty downstream decisions.

We demonstrate a novel unsupervised algorithm for detecting whether or not a given input to a deployed model is anomalous, without requiring the input to be labeled. Further, a practical approach for discovering new emerging classes in the edge environment is showcased. The algorithm is evaluated on the MNIST character recognition dataset. The results show that the anomaly detection algorithm catches 93% of the anomalies when the model is deployed in a new domain. Further, only 1.5% of the normal inputs are falsely identified as anomalies. Via existing clustering techniques, we also show the effectiveness of our algorithm in discovering new classes from the anomalous inputs.


Training subset,
One subset per class label
Labeled
Training data
Model
Layer N
output
Class-wise
centroids,
normalized
inertia
Edge
deployment
Centroids /
inertia
Cloud
Edge
distance >
threshold
Anomalous
inputs
Relative distance
of layer N output
with centroids
Open set
Input, no label
distance <
threshold
Use for prediction