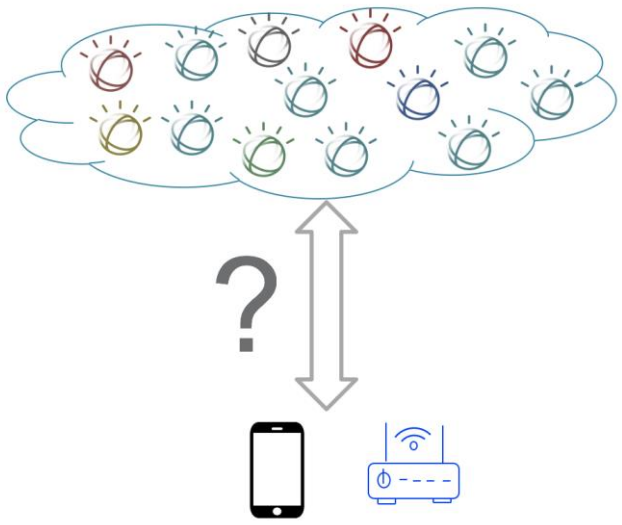
Unsupervised estimation of domain applicability of models

Nirmit Desai, Linsong Chu, Raghu K. Ganti, Heesung Kwon, Ian Taylor, Mudhakar Srivatsa

Problem: Domain Match Estimation



- For a given unlabeled input and a model, how likely is it that the input is “outside” the domain of the model?
- For a set of unlabeled inputs and a model repository, how do we select the most suitable model?
- For a given unlabeled input and a model, should we use the input for prediction OR skip the input given that it may lead to an erroneous prediction?

Approach: Investigate metrics for domain similarity

$$I(C,X) = \sum_{i=1}^{N} \frac{(C^* - X_i)^2}{N}$$

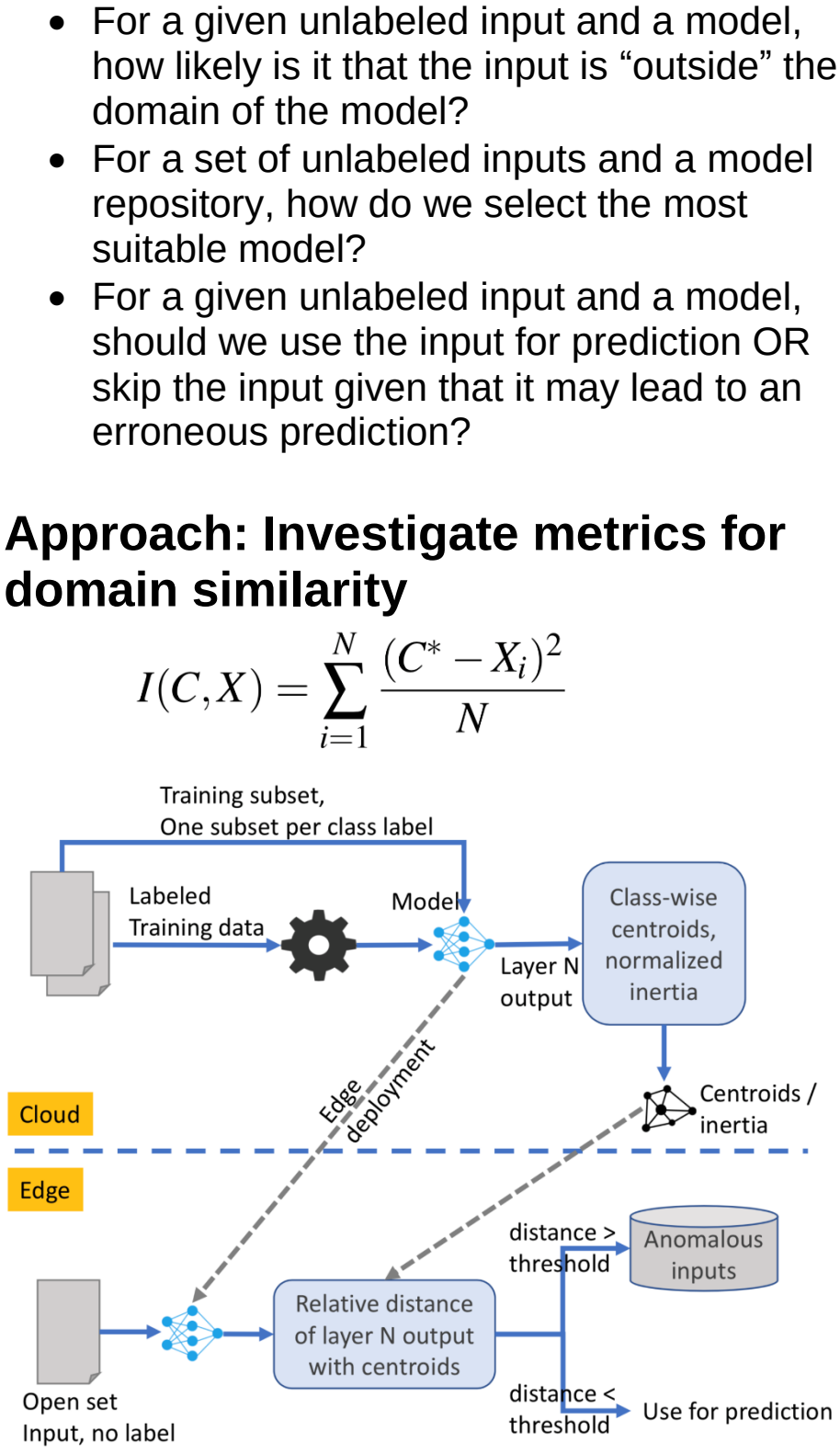


Fig 1: System Architecture

Results: MNIST dataset

- I_t → Inertia computed on training data
- I → Inertia computed on unlabeled inputs

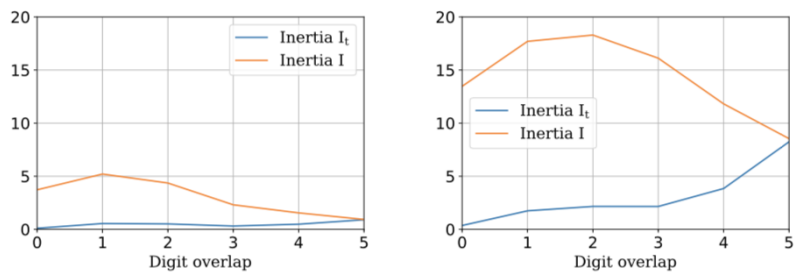


Figure 2: Left: Inertial on layer11, Right: Inertia on layer12 (domain1: 0-4, domain2: 5-9)

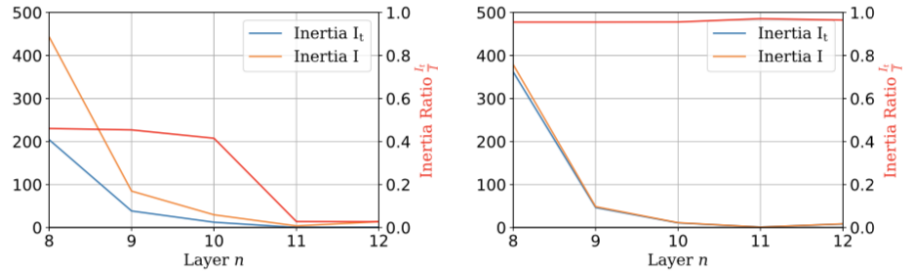


Figure 3: Left: no overlap between domain1 and domain2, Right: five digit overlap

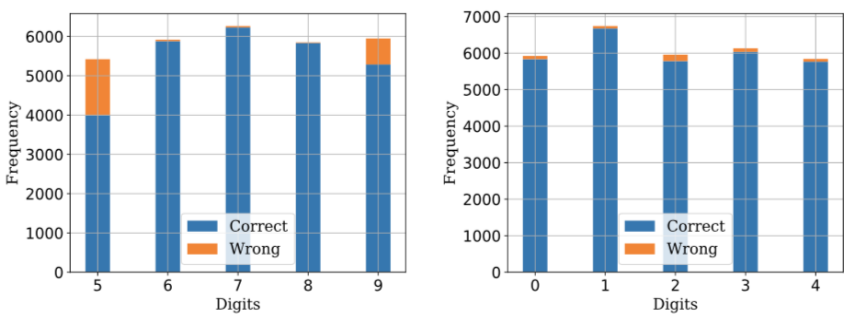


Fig 4: Anomaly detection accuracy: 92% of true positives, 1.5% false positives (Left: domain2 inputs, Right: training data)

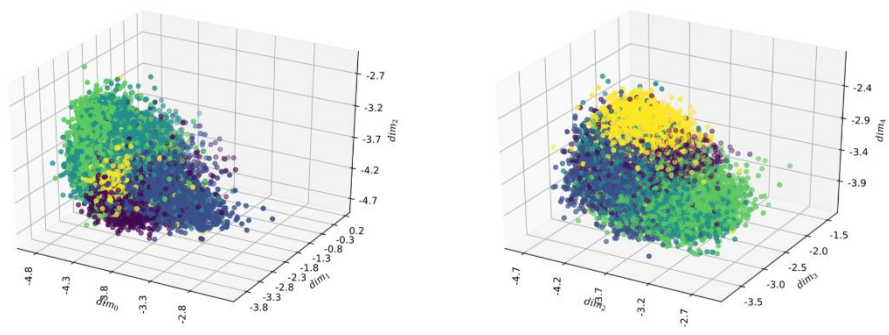


Fig 5: Layer12 outputs of domain2 inputs suggesting emergence of new classes

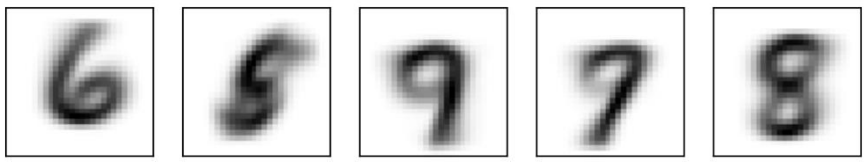


Fig 6: Centroids of layer12 outputs on domain2 raw inputs