Message Passing Adaptive Resonance Theory for Online Active Semi-supervised Learning

Taehyeong Kim[1,2] *Injune Hwang*[1] *Hyundo Lee*[2] *Hyunseo Kim*[2]
Won-Seok Choi[2] *Joseph J. Lim*[1,3] *Byoung-Tak Zhang*[2]

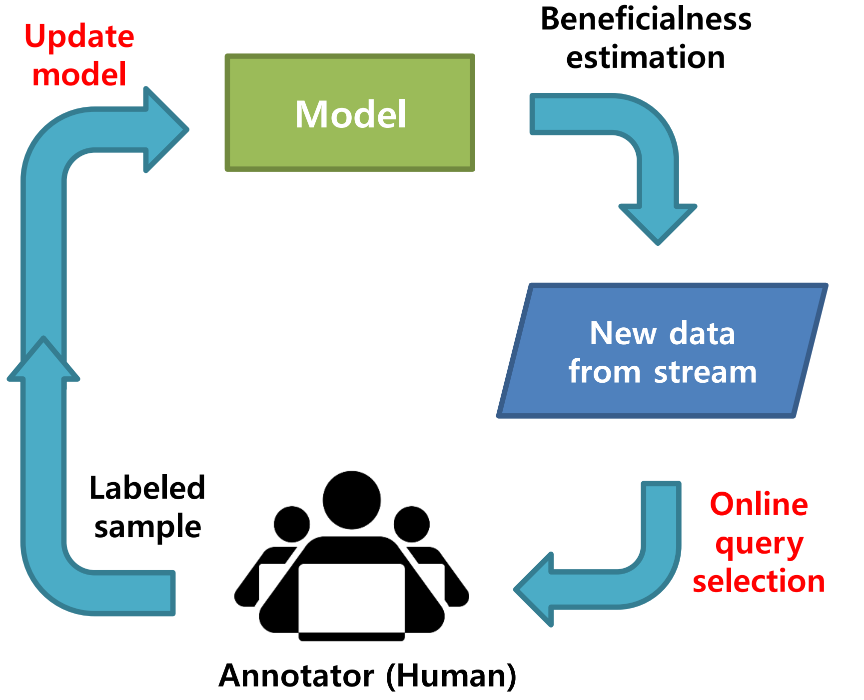
[1]AI Lab, CTO Division, LG Electronics, Seoul, Republic of Korea
[2]Seoul National University, Seoul, Republic of Korea
[3]University of Southern California, California, USA

The Real-world Problems

- The recent success of deep learning is largely attributed to massive amount of labeled data.

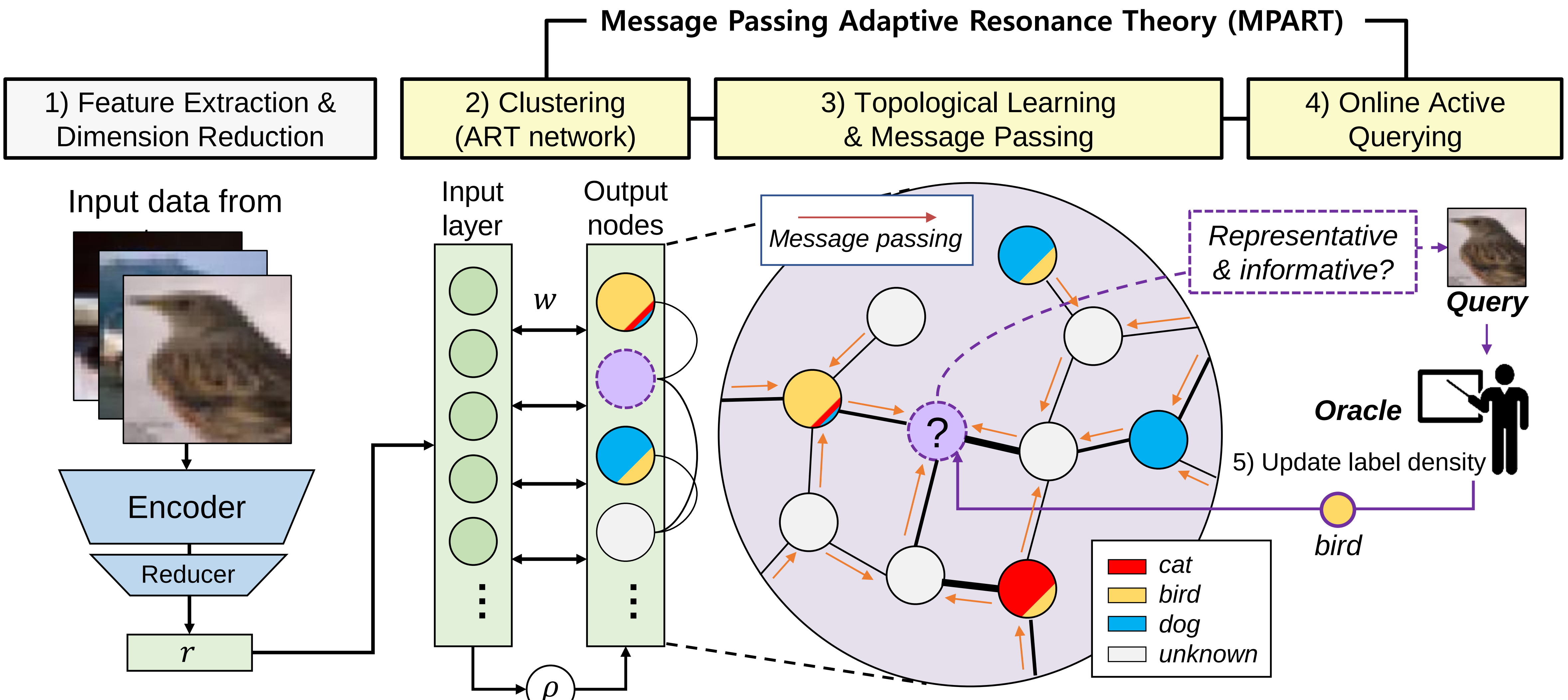
- **However, the real-world problems**..
  - Most of the data samples are **not labeled**.
  - The **collection and storage** of personal data are **prohibited** due to privacy concerns.

- This burden impedes the **widespread** use of deep learning in **real-world applications**.



<Stream-based Online Active Learning>

Model Overview

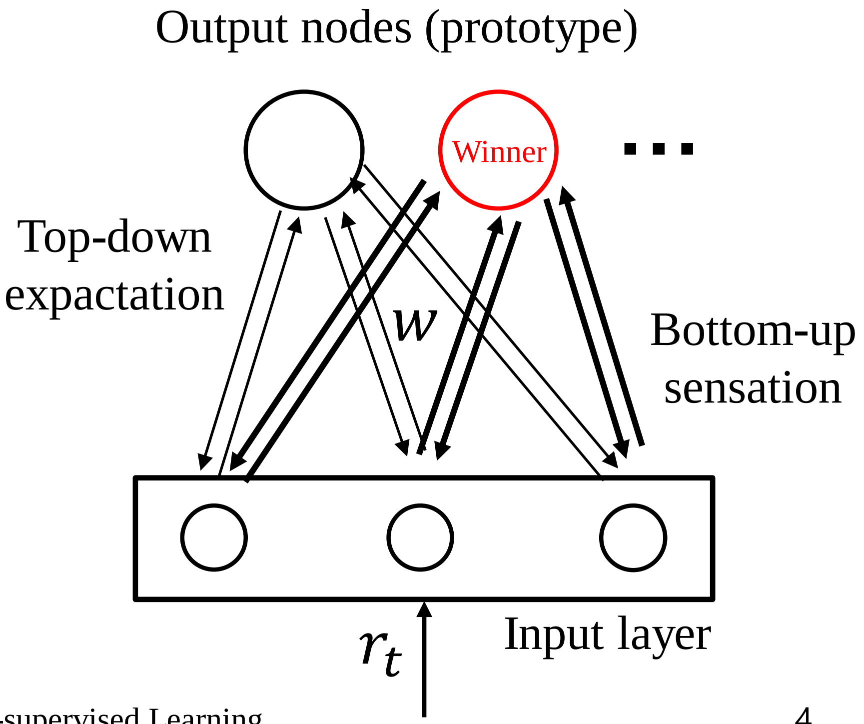


ART Networks

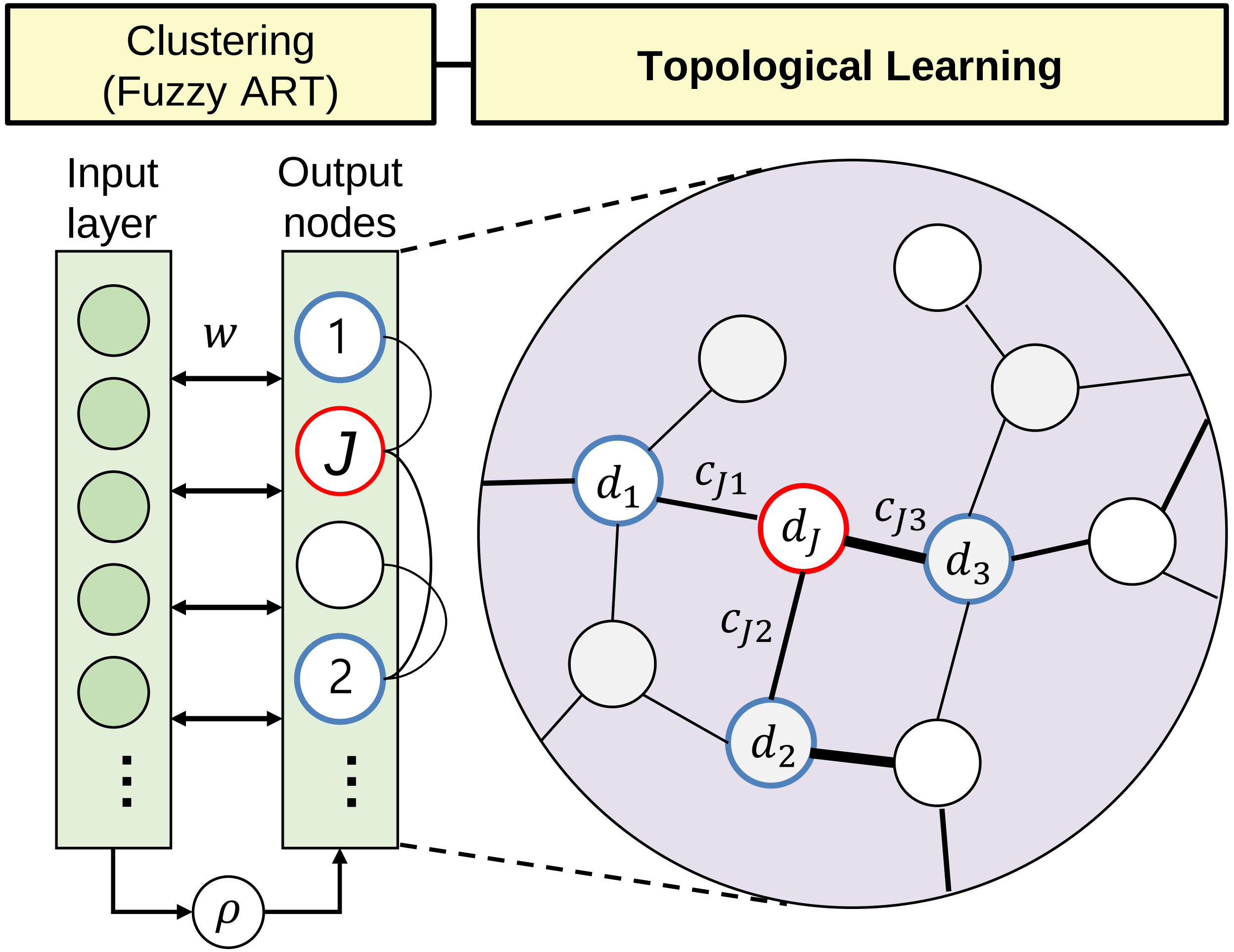
- Adaptive Resonance Theory (ART), inspired by brain information processing mechanisms, is an **unsupervised learning method** for pattern recognition.

- In terms of **being conservative while learning new**, the ART networks can be **a solution for the online learning**.

- We use Fuzzy ART to **form nodes in a topological graph** through clustering of input data.



Topology Learning with MPART



J Winner node

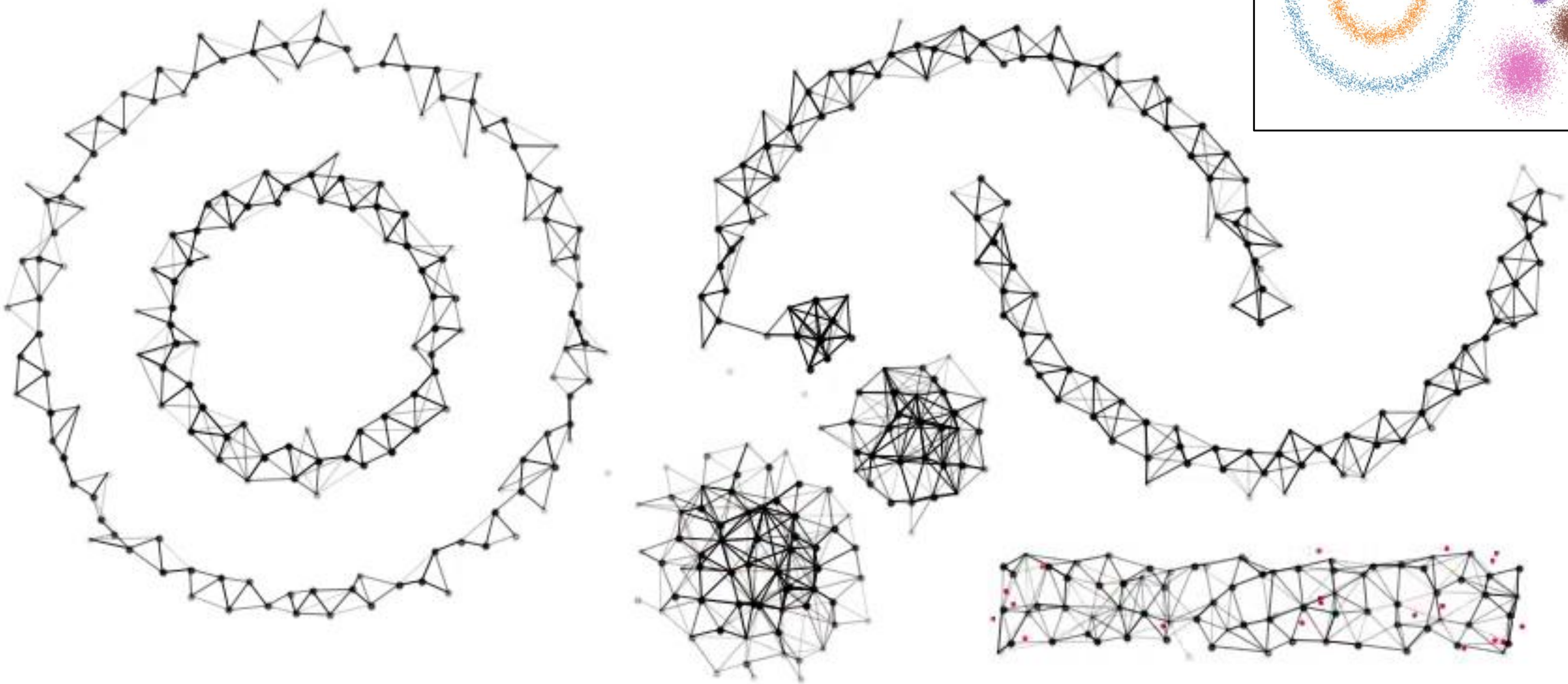
Co-activated nodes

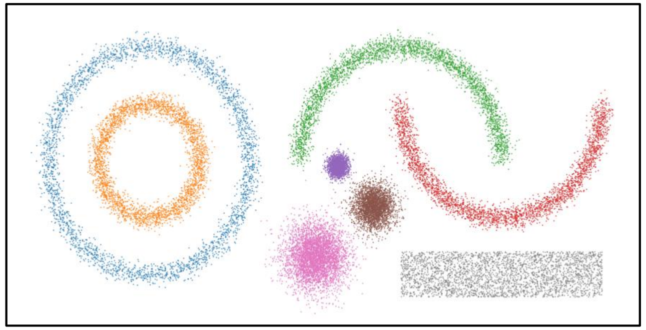
$$d_i, d_j = winning\ count\ of\ nodes$$

$$c_{ij} = co\text{-}activation\ count$$

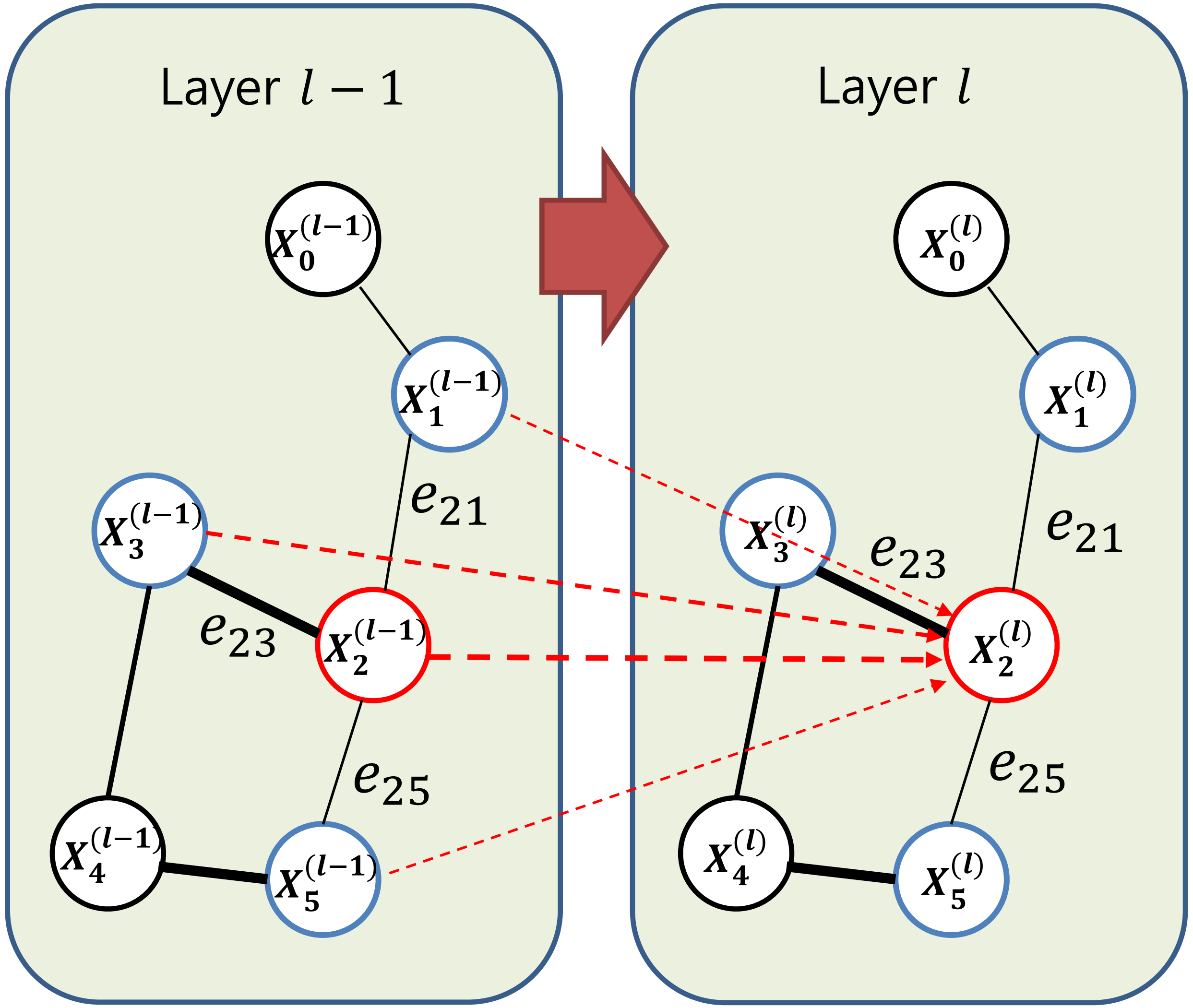
$$e_{ij} = edge\ weight\ between\ node\ i\ and\ j$$
$$= \frac{c_{ij}}{d_i + d_j}$$

Topology Learning with MPART





Message Passing



$$X_i^{(l)} = X_i^{(l-1)} + \delta \sum_{j \in \mathcal{N}_i^{(1)}} e_{ij} X_j^{(l-1)}$$

$X_i^{(l)} = an\ information\ vector$

$\mathcal{N}_i^{(l)} = a\ set\ of\ l\text{-}hop\ neighbors$

$\mathcal{N}_i^{(0:l)} = a\ union\ of\ \mathcal{N}_i^{(0)}, \dots, \mathcal{N}_i^{(l)}$

$e_{ij} = edge\ weight$

$\delta = propagation\ rate$

Node Classification

- **The class information $\boldsymbol{q_i}$** of a node i, called the ***label density***, is a distribution over a set of known class labels C.
  - Indicates how probable the node belongs to each class.
  - When a label y_t is provided with an input x_t, the corresponding density value in the winner J_t, i.e. $q_{J_t}(y_t)$, increases by 1.

- Infers the class of input x_t by **aggregating the class information of neighboring nodes.**
  - We obtain the class probability distribution $\boldsymbol{p_t}$ of winner node J_t by performing ***L***-layer message passing and normalizing it as follows.

$$p_t(y) = q_{J_t}^{(L)}(y) / \sum_{y' \in C} q_{J_t}^{(L)}(y')$$

Density-Weighted Uncertainty Estimation

- **We use two kinds of uncertainty measures for nodes.**
  - Combine them to get a query selection score.

$$u_e = 1 - \tanh\left(k_e \sum_{y \in C} q_{J_t}^{(L)}(y)\right) \qquad u_a = \begin{cases} \dfrac{-\sum_{y \in C} p_t(y) \log(p_t(y))}{\log(|C|)}, & if\ |C| > 1 \\ 0, & otherwise \end{cases}$$

$$u_t = \tau \cdot u_\mathrm{e} + (1 - \tau) \cdot u_a$$

- Finally, the **density-weighted query selection score** s_t using **distribution density** $d_{J_t}^{(L)}$ is defined as follows.

$$d_i^{(l)} = d_i^{(l-1)} + \delta \sum_{j \in \mathcal{N}_i} e_{ij} d_j^{(l-1)}$$

$$s_t = \tanh\left(k_d \cdot d_{J_t}^{(L)}\right) \cdot u_t$$

Tasks

- Online active learning task for **stream-based selective sampling**.
  - **Multi-class classification**, where the **number of classes is not known**.
  - The number of queries is limited to a **fixed budget *B*** within a **period of *W*** consecutive inputs. (Query frequency = ***B/W****)*

- Experimental settings
  - **Four kinds of datasets with different distributions**: Mouse retina transcriptomes, Fashion MNIST, EMNIST Letters, and CIFAR-10.
  - Query frequency: ***B* = 1 or 2**, **W = 100, 500, 1,000 or 2,000 (e.g. 1/1000, 1/500)**.
  - We only used **10,000 randomly sampled data from the training split per trial**.
    → The total query budget is 10~100.

Query Selection Strategy

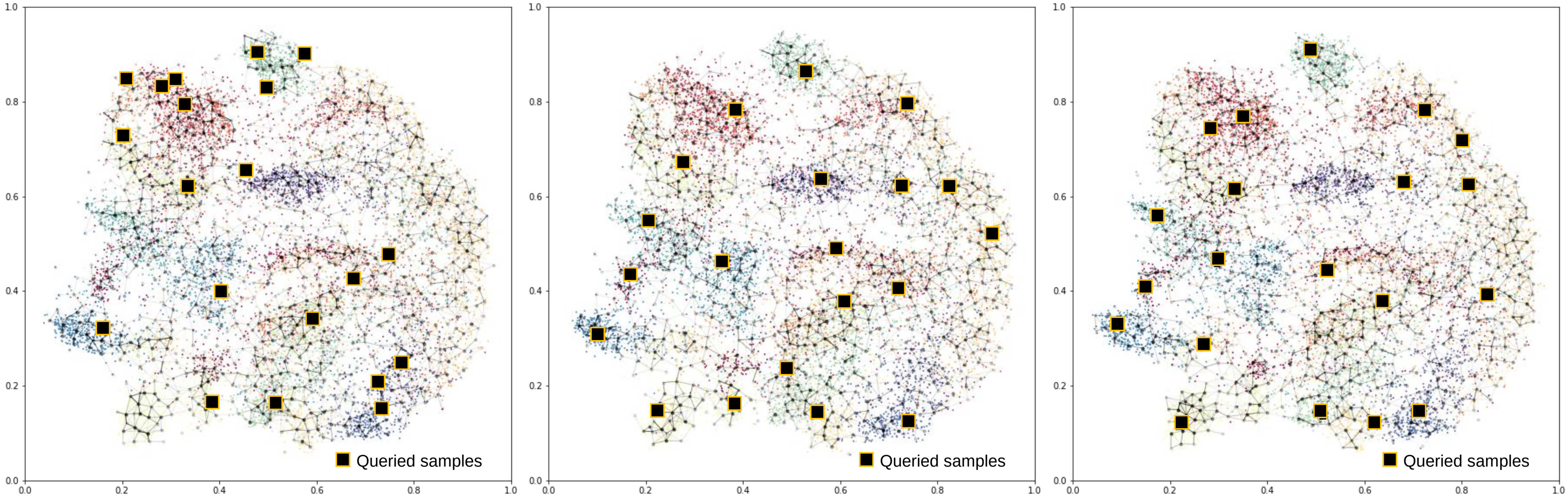
- **Random** : A random query is selected from a sequence of inputs.

- **Memory** : The model has a memory that can store one sample.
  - One sample with **the maximum query selection score** is stored.
  - The stored sample is queried at the end of the query period.

- **Explorer** : The model **cannot store any input sample**.
  - **Selects *B* samples online** for each query period *W*.
  - **The uncertainty distribution** of input data is continuously **estimated**.
  - Useful samples are selected by predicting the beneficialness.

Experimental Results

- Visualization of training results on EMNIST Letters
  - The number of layer $L = 3$ and $B/W = 1/1000$ were used.



(a) Random (b) Memory (c) Explorer

Conclusions

- We propose **Message Passing Adaptive Resonance Theory (MPART)** for online active semi-supervised learning.

- MPART **learns the distribution and the topology of the input data online**, infers the class of unlabeled data, and **selects the informative and representative samples through message passing** between nodes on the topological graph.

- We believe MPART offers new opportunities for machine learning techniques to be widely used in real-world applications.

Taehyeong Kim

thkim@bi.snu.ac.kr