Continual Self-organized Learning of Hierarchical Multimodal ART
Il-Jae Kwon\textsuperscript{1,}\textsuperscript{2}, Byoung-Tak Zhang\textsuperscript{1,}\textsuperscript{2,}\textsuperscript{3}
Brain Science Program\textsuperscript{1,}, Department of Computer Science and Engineering\textsuperscript{2}, Program in Cognitive Science\textsuperscript{3}
Seoul National University

Abstract
When people learn a concept of an object, we rely on many different sensory data. Depending on one single sensory system can possibly cause some confusion, but conjugating multiple systems in learning can resolve such uncertainties. In addition, human learning effectively utilises incremental learning strategies, which means that learning process continues as new training samples emerge. Taking advantage of these two observations, we suggest the Hierarchical Multi-modal Adaptive Resonance Theory model (HiMART). HiMART, in addition to taking advantage of incremental learning, builds the generalized multi-modal patterns that can later be used as long-term memories. Finally, we discuss further experiments using our model.

1 Introduction
Despite the great success in mainstream machine learning based on gradient descent, such mainstream methods have many drawbacks. Training a model requires a tremendous amount of data and is not suitable for continual learning [11]. However, human learning is often represented as one-shot learning and incremental learning [12]. Even one example can be stored permanently in the human brain as a learned concept [13] and further generalize to other new examples. To generalize and create more concepts, humans use incremental learning. In this sense, Adaptive Resonance Theory [14, 5], also called as ART, is a good solution for modeling human learning, capable of both one-shot learning and incremental learning. Another striking finding in human learning is the Grandmother Cell, also referred to as the Jennifer Aniston cell, which is a hypothetical neuron that fires not only when one is looking at a picture of their grandmother but also when one reads his grandmother’s name. [6] found that these kinds of cells actually exist in the human brain. They found that such Grandmother Cells are activated by the multi-modal concept of the grandmother. Grandmother Cell exists in the medial temporal lobe, which is a region that is essential for declarative memory and operates in conjunction with neocortex to build long-term memories [7]. This implies that some long-term memories are actually invariant to input systems and act as a complex cell that is activated by several other modalities. In addition, [8] found out that hippocampus in medial temporal lobe is crucial for relational memories. Relational memories, in the context of neuroscience, resembles relations of memories such as co-presence, co-occurrence [9], higher-level relationships and cause-effect relationships. In this paper, we propose a method to build relations between long-term memories.

2 Preliminaries
2.1 Adaptive Resonance Theory
Elementary ART is composed with two different fields: a comparison field and a recognition field. Comparison field is the layer that holds all long-term patterns $P_j$. Recognition field, also referred to as the feature representation field, is where input signal $x$ is forwarded into bottom-up activation function $F_1$. The resulting $T_j$ will be a metric on how much input $x$ fits to pattern $P_j$.

$$T_j = F_1(x, \theta_j^{up})$$ (1)

where $\theta_j^{up}$ is a bottom-up weight for pattern $P_j$. After computing all similarity measures $T_j$ for every $j$, we find the best matching pattern $j$ that maximizes the similarity measure. Once the best matching pattern is decided, we verify if the selected pattern is good enough, by computing confidence score using a top-down function $F_2$. Confidence score $c$ is higher if the input better fits the pattern.

$$c_j = F_2(x, \theta_j^{down})$$ (2)

If the confidence score $c_j$ of the pattern $j$ is above the threshold parameter $\rho$, the network accepts the example to the pattern and the pattern will enter a resonance state. In the resonance state the weights $\theta_j^{up}$ and $\theta_j^{down}$ will be updated using a learning function of accepted sample $x$ and learning rate $\beta$. However, if no other pattern can pass the verification test, a new pattern is generated with example $x$ as it’s first sample. ART is inspired by how the brain processes and generates patterns of examples from an ever changing environment. ART is known as a good tool for incremental learning because it preserves the existing patterns even when there is a new input. Taking advantage of the stable and incremental learning ability of ART, we present the Hierarchical Multi-modal Adaptive Resonance Theory model or HiMART.

3 Methods
HiMART is composed with hierarchically stacked ART modules. It has two levels where the bottom level is composed of separate ART modules for each modal system; for example in Figure 1, system 1 exploits the vision feature, while system 2 exploits the human auditory feature. The upper level ART, the Abstract ART, will receive concatenated patterns that are from each bottom level ARTs, the System ARTs, and determine the multi-modal abstract patterns that can serve as long-term memories with relations.

3.1 Hierarchical ART
Starting with sets of features from different modalities, examples are fed to the separate System ARTs corresponding to their modalities. All sets of inputs are encoded before entering ART by adequate encoders, such as pre-trained convolutional neural network, according to their data types. Each System ART module will output different system patterns according to the similarities between the examples they see. However, unlike original ART, we will keep $N$ accepted samples that received top-$N$ confidence score and use a mean aggregator or max aggregator to create a single representation vector for each system pattern. $N$ samples are limited for
as human speech input like “This is a dog” is added to those two visual patterns, the similarity between “Poodle” and “Golden Retriever” will increase as both of these dog species will be related to the “this is a dog” audio pattern. Thus, the model is able to create an abstracted and generalized multi-modal patterns.

3.2 Relations between the Abstract Patterns

\[
\text{Score}_{kl} = \sum_{i \in S} C_{ik} \times C_{il}, \quad w_{kl} = \frac{\text{Score}_{kl}}{\sum_k \sum_l \text{Score}_{kl}} (3)
\]

For now, we assume that every systems has equal importance, but adapting different weights of importance among systems would be an interesting avenue for future research. Simply multiplying different constants to the relation scores based on the systems is

![Hierarchical Adaptive Resonance Theory Model](image)

To illustrate, if there are three systems as shown in Figure 2, an example set will first be encoded corresponding to each of their systems. Each encoded features of the example set will be fed to a System ART and create system pattern \(S_i^k\). Once all of the \(S_i^k\) from the example set are determined, we concatenate them. Concatenated \(S_i^k\) will now be used as a new input for the upper level ART, an Abstract ART. Abstract ART is system-independent. By using concatenated patterns as a new input, we gain three advantages. The first advantage is that our model can continually make multi-modal patterns using multiple systems. Secondly, the resulting abstract patterns \(P_k\) that can later be used as long-term memories.

![Making each pattern representation from the outputs of System ART modules](image)

![Making each pattern representation from the outputs of System ART modules](image)
4 Experimental Setup

4.1 Multi-modal Incremental Learning

Our model, HiMART, can incrementally learn multi-modal samples. Abstract patterns will be updated to form generalized long-term memories which can be used for multi-modal classification tasks. We can define the memory from the best matching abstract pattern and related patterns from topological relations. With $P_i$ representing the best matching abstract pattern for input $x$,

$$\text{memory} = P_i + \sum_{j \neq i} w_{ij} \times P_j \quad (4)$$

Then along with the input $x$, we will concatenate the memory with input, then feed the resulting vector to a fully-connected layer to train a multi-modal classifier.

4.2 Multi-Label Zero-shot Learning

Another task to utilize HiMART is zero-shot learning. Zero-shot learning is an ability to label the examples of an unseen class. Multi-label zero-shot learning is a task which shows pictures containing multiple objects (containing unseen labels) and requires the model to label all the detected objects in the image. Previous works on zero-shot learning mostly relies on the language relations generated from text corpus [[10]]. However, our model generates it’s own topological relations that can also be used as an auxiliary knowledge for zero-shot learning. This task is not strictly a multi-modal task, but it proves the benefits of the relations between the patterns generated by HiMART. HiMART itself acts as a classifier without extra fully-connected layers. In this task, system1 will be the whole image containing all objects and system2 will be the cropped image of a single detected object. By having the whole image in system1, the HiMART will give higher relation score if two objects were frequently shown together in the same image. By labeling the resulting abstract pattern with the label of the detected object, we can decide which label is the majority among $N$ samples in the abstract pattern. Using the relation between patterns (labels) from HiMART that we discussed in Section 3.2, on top of word relations from text corpus, we can attempt zero-shot learning. Relations between words can be expressed in many ways. [11] utilized WordNet [12] and classified the relation between words into three categories: super-subordinate, positive, negative. A positive relation is a word pair that has similar semantics, while a negative relation is a word pair with antithetical semantic meaning that should be avoided. After sufficient examples are fed into HiMART, we use the labels of the patterns to create a graph where labels become the nodes. Word relations from WordNet and pattern relations $w_{ij}$ from HiMART are the edges of the graph. Each node in the graph will have a node state. After going through several layers of Graph Convolutional Network (GCN) [13], each of the resulting node states will be fed to a fully-connected layer for binary-classification: exist or not exist. Once the binary-classifier is trained, we add the unseen labels as a node and connect nodes based on the word relations from WordNet. Once again, we use GCN to diffuse the information between the seen labels and the unseen labels. Using the latest node state, we feed it to the trained binary-classifier to decide if that label exists or not.

4.3 Acknowledgement

This work was partly supported by the Institute for Information Communications Technology Promotion (2015-0-00310-SW,StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind) and Korea Institute for Advancement Technology (P0006720-GENKO) grant funded by the Korea government.