공학박사학위논문

동적 멀티모달 데이터 학습을 위한 심층 하이퍼네트워크

Deep Hypernetworks for Learning from Nonstationary Multimodal Data

2015년 2월

서울대학교 대학원

전기컴퓨터공학부

하 정 우

Abstract

Recent advancements in information communication technology has led the explosive increase of data. Dissimilar to traditional data which are structured and unimodal, in particular, the characteristics of recent data generated from dynamic environments are summarized as high-dimensionality, multimodality, and structurelessness as well as huge-scale size. The learning from non-stationary multimodal data is essential for solving many difficult problems in artificial intelligence. However, despite many successful reports, existing machine learning methods have mainly focused on solving practical problems represented by large-scaled but static databases, such as image classification, tagging, and retrieval.

Hypernetworks are a probabilistic graphical model representing empirical distribution, using a hypergraph structure that is a large collection of many hyperedges encoding the associations among variables. This representation allows the model to be suitable for characterizing the complex relationships between features with a population of building blocks. However, since a hypernetwork is represented by a huge combinatorial feature space, the model requires a large number of hyperedges for handling the multimodal large-scale data and thus faces the scalability problem.

In this dissertation, we propose a deep architecture of hypernetworks for dealing with the scalability issue for learning from multimodal data with non-stationary properties such as videos, i.e., deep hypernetworks. Deep hypernetworks handle the issues through the abstraction at multiple levels using a hierarchy of multiple hypergraphs. We use a stochastic method based on Monte-Carlo simulation, a graph MC, for efficiently constructing hypergraphs representing the empirical distribution of the observed data. The structure of a deep hypernetwork continuously changes as the learning proceeds, and this flexibility is contrasted to other deep learning models. The proposed model incrementally learns from the data, thus handling the nonstationary properties such as concept drift. The abstract representations in the learned models play roles of multimodal knowledge on data, which are used for the content-aware crossmodal transformation including vision-language conversion. We view the vision-language conversion as a machine translation, and thus formulate the vision-language translation in terms of the statistical machine translation. Since the knowledge on the video stories are used for translation, we call this story-aware vision-language translation.

We evaluate deep hypernetworks on large-scale vision-language multimodal data including benmarking datasets and cartoon video series. The experimental results show the deep hypernetworks effectively represent visual-linguistic information abstracted at multiple levels of the data contents as well as the associations between vision and language. We explain how the introduction of a hierarchy deals with the scalability and non-stationary properties. In addition, we present the story-aware vision-language translation on cartoon videos by generating scene images from sentences and descriptive subtitles from scene images. Furthermore, we discuss the meaning of our model for lifelong learning and the improvement direction for achieving human-level artificial intelligence.

Keywords: **Deep hypernetwork, Higher-order graphical model, Nonstationary multimodal data, Multimodal concept learning, Stochastic hypergraph construction, Incremental learning, Vision-language translation**

Student Number: 2006-21317

Contents

Abstract **i**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Problems to be Addressed 3

1.3 The Proposed Approach and its Contribution 3

1.4 Organization of the Dissertation 6

2 Related Work **9**

2.1 Multimodal Leanring 9

2.2 Models for Learning from Multimodal Data 10

2.2.1 Topic Model-Based Multimodal Leanring 11

2.2.2 Deep Network-based Multimodal Leanring 13

2.3 Higher-Order Graphical Models 15

2.3.1 Hypernetwork Models 16

2.3.2 Bayesian Evolutionary Learning of Hypernetworks 19

3 Multimodal Hypernetworks for Text-to-Image Retrievals **23**

3.1 Overview 23

3.2 Hypernetworks for Multimodal Associations 26

3.2.1 Multimodal Hypernetworks 26
3.2.2 Incremental Learning of Multimodal Hypernetworks 28
3.3 Text-to-Image Crossmodal Inference 29
3.3.1 Representatation of Textual-Visual Data 30
3.3.2 Text-to-Image Query Expansion 30
3.4 Text-to-Image Retrieval via Multimodal Hypernetworks 33
3.4.1 Data and Experimental Settings 33
3.4.2 Text-to-Image Retrieval Performance 34
3.4.3 Incremental Learning for Text-to-Image Retrieval 38
3.5 Summary 40

4 Deep Hypernetworks for Multimodal Cocnept Learning from Cartoon Videos 41
4.1 Overview 41
4.2 Visual-Linguistic Concept Representation of Catoon Videos 44
4.3 Deep Hypernetworks for Modeling Visual-Linguistic Concepts . . . 44
4.3.1 Sparse Population Coding 46
4.3.2 Deep Hypernetworks for Concept Hierarchies 48
4.3.3 Implication of Deep Hypernetworks on Cognitive Modeling . 52
4.4 Learning of Deep Hypernetworks 53
4.4.1 Problem Space of Deep Hypernetworks 53
4.4.2 Graph Monte-Carlo Simulation 54
4.4.3 Learning of Concept Layers 57
4.4.4 Incremental Concept Construction 58
4.5 Incremental Concept Construction from Catoon Videos 60
4.5.1 Data Description and Parameter Setup 60
4.5.2 Concept Representation and Development 62

4.5.3 Character Classification via Concept Learning 65
4.5.4 Vision-Language Conversion via Concept Learning 66
4.6 Summary 68

5 Story-aware Vision-Language Translation using Deep Concept Hiearachies 69
5.1 Overview 69
5.2 Vision-Language Conversion as a Machine Translation 73
5.2.1 Statistical Machine Translation 73
5.2.2 Vision-Language Translation 75
5.3 Story-aware Vision-Language Translation using Deep Concept Hierarchies 77
5.3.1 Story-aware Vision-Language Translation 77
5.3.2 Vision-to-Language Translation 80
5.3.3 Language-to-Vision Translation 81
5.4 Story-aware Vision-Language Translation on Catoon Videos 82
5.4.1 Data and Experimental Setting 82
5.4.2 Scene-to-Sentence Generation 84
5.4.3 Sentence-to-Scene Generation 87
5.4.4 Visual-Linguistic Story Summarization of Cartoon Videos . . 90
5.5 Summary 91

6 Concluding Remarks 94
6.1 Summary of the Dissertation 94
6.2 Directions for Further Research 96

Bibliography 98

List of Tables

3.1 Parameter setup for model learning 34
3.2 Contents of two types of textual queries 35
3.3 Precision and recall of text-to-image retrieval for 10 queries 35
3.4 Successful retrieval for 30 queries 36

4.1 Comparison of deep hypernetworks to other multimodal leaning models 52
4.2 Character classification results 66

5.1 Scene-to-sentence translation performance on the *Pororo* dataset . . . 85
5.2 Scene-to-sentence translation performance on Flickr.com dataset . . . 87

List of Figures

1.1 Paradigm shift from static unimodal data to non-stationary multimodal data 2

1.2 Improvement of the proposed models in this dissertation 5

1.3 Main results and used data for the proposed models 6

2.1 Graphical representation of the LDA model (a) and the corr-LDA model (b). Gray boxes denote observable variables. 11

2.2 Graphical representation of the multimodal deep Boltzmann machine (a) and the multimodal neural language model (b) 14

2.3 An example of a hypernetwork and its term 19

2.4 Overall procedure for learning Bayesian evolutionary hypernetworks. (a) presents the flow of learning hypernetworks with a functional level. (b) explains the evolution of hypernetworks from an Bayesian point of view. 20

3.1 Overall flow of text-to-image retrieval using incrementally learned multimodal hypernetworks 25

3.2 Hyperedges consisting of three texutual words and two visual pateches from a textual-visual data instance (a captioned image) 27

3.3 Flow of constructing visual patches from a given image with MSER and SIFT, two feature extraction methods. Boxes and circles denote the external regions extracted by MSER and the salient features by SIFT, respectively 31

3.4 Flow of the crossmodal query expansion from the given textual query to the visual query 32

3.5 Precision and recall of text-to-image retrieval using 1000-hyperentworks for 30 queries on (a) the training dataset and (b) the test set. Values are averaged for 30 queries 36

3.6 Retrieved images from the test dataset using 1000-hypernetworks for three queries. Red boxes are images including query-objects 37

3.7 Precision and recall of text-to-image retrieval using 1000-hypernetworks for 10 queries on the test set as the retrieval size grows 38

3.8 (a) Vocabulary size of hyperentworks and (b) average performance of text-to-image retrieval for 10 queries as learning proceeds incrementally. In (b), the number of retrieved images is 20. 39

3.9 Retrieved images on the test dataset for two textual queries such as 'sky' and 'grass' with 300-hypernetworks and 1000-hypernetworks. Red boxed photos are images including the query-object. 40

4.1 Examples of utterance-scene pairs extracted from cartoon videos . . 45

4.2 Architecture of deep hypernetwork for representing concept hierarchies 46

4.3 An example of the concept representation of Pororo via sparse population coding. SC denotes a sub-concept corresponding to a microcode 48

4.4 An example of sparse population coding model and its hypernetwork representation 49

4.5 An Example of deep concept hierarchy with two conceptual layers learning from cartoon videos. Abstract concept variables denote the characters appearing in the videos. Gray boxes are observable variables. 50

4.6 Visual-linguistic representation and development of three character concepts of video contents. A scene-utterance pair is represented by the sets of image patches and words and the concepts of the video stories are represented by these patches and words. *Tongtong* is not seen in episodes 1-13 and appears in episode 56 for the first time. . . 61

4.7 Evolution of *Pororo*'s visual-lingustic concept maps 62

4.8 Changes of model complexity according to the learning strategies of the graph MC. In (b), VU, VPR, and VF denote the vertex sets of the model constructed by UGMC, PRGMC, and FGMC 63

4.9 PCA plot of microcodes associated with the concrete concept nodes (c1-nodes) and their centroids of the models learned from 183 episodes by UGMC. 64

4.10 Results of vision-language conversion. precision of vision-to-language (a) and intermediate images generated from given sentences (b) . . . 67

5.1 Example of sentences generated by conventional methods and story-aware vision-language translation. 78

5.2 Vision-language translation by the crossmodal inference via a deep hypernetwork 79

5.3 Story-aware subtitle generation for given scene images as the increase of the observed videos. SPC denotes a model with no concept layer. 83

5.4 Precision (a) and recall (b) of scene to sentence as the increase of the observed videos generation 84

5.5 Examples of sentences (words) generated by DCHs and other models 86

5.6 Generated intermediate images from given sentences as the increase of the observed videos 88

5.7 Changes of the relationships between main characters as the story proceeds 89

5.8 Visual-linguistic video summarization based on the changes of character relationships 90

Chapter 1

Introduction

1.1 Background and Motivation

The progress of information technology has rapidly increased the complexity as well as the quantity of data. The paradigm of representing and deliverying information shifts from single modality to multimodal representations as shown in Figure 1.1. A great many people use internet and social network services using PCs, tablet PCs, and smart phones and their behaviors and thoughts are stored as life-logs. Unlike conventional data that are relatively low-dimensional, structured, numerical and unimodal, these data recently generated including videos, images, and social network service logs can be summarized in terms of high-dimensionality, structurelessness, and multimodality. In addition, they are continuously generated from not static and stationary but dynamic and non-stationary environments. In particular, the data generated by human behaviors are very helpful for studying and modeling human beings. Therefore, the use of these complex multimodal data from dynamic environments is essential for achieving human-level artificial intelligence (Muggleton, 2014).

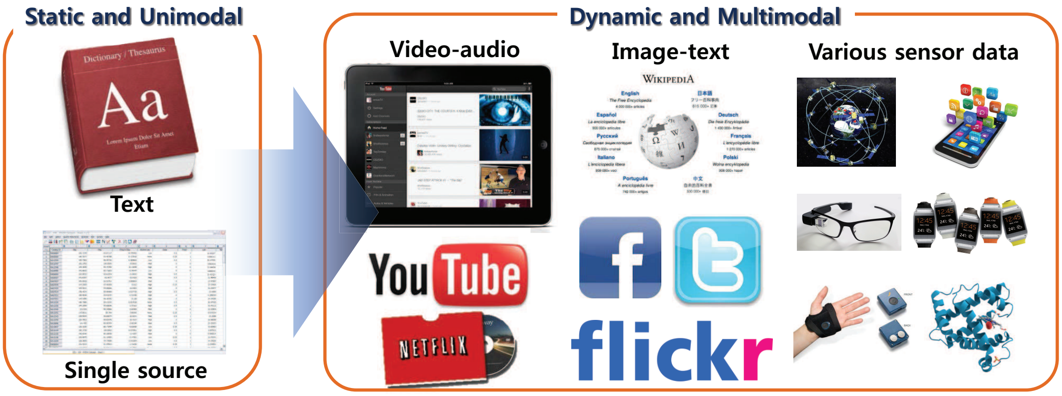


Figure 1.1: Paradigm shift from static unimodal data to non-stationary multimodal data

Machine learning has been successfully applied to various real-world applications for past three decades and recent advancement in machine learning have led the great advancement in data mining and artificial intelligence. Emergence of milestone methods such as Bayesian networks, support vector machines and deep neural networks has helped other fields including industry, politics, natural and social sciences to advance. However, many existing machine learning models mainly focused on solving specific problems rather than developing generalized solution such as knowledge construction. Moreover, the data on which the traditional methods are evaluated are mostly generated from static environments. Even though incremental methods were introduced for learning from large-scale data, many methods rarely consider the pattern changes implicated in the data as the progress of time factors, i.e., concept drift, with multiple time frames. The learning in short and long term resolutions is necessary for implementing lifelong learning (Zhang, 2013) to reach human-level intelligences systems.

1.2 Problems to be Addressed

Mutimodal learning is a method for finding the association rules by learning from multimodal data. The association rules can be used for the knowledge on the data as well as various applications such as image retrievals, image annotation, and description generation. Since the semantics of the data are mainly represented by the associations among more than three multimodal features including words, phonemes and image patches rather than pair-wise feature relations, higher-order models are suitable for characterizing the data contents. In addition, learning from large-scale non-stationary data such as videos requires incremental methods which can handle concept drifts implicated in the contents.

This dissertation proposes a higher-order probabilistic graphical model i.e., a hypernetwork for learning from non-stationary multimodal data and focuses on incrementally constructing and representing knowledge on the data while dealing with concept drifts. Learning non-stationary multimodal data with hypernetworks involves three technical issues as follows:

i) representing higher-order associations among multimodal features,

ii) exploring the huge combinatorial space representing a hypernetwork,

iii) learning from continuously increasing data while handling concept drifts.

In this dissertation, we used many numbers of the series of cartoon videos as non-stationary multimodal data and addressed these three issues in learning the model from the cartoon video data.

1.3 The Proposed Approach and its Contribution

We propose an advanced hypernetwork for learning the associations among visual-linguistic features from large-scale non-stationary multimodal data and representing multiple levels of multimodal features by introducing a hierarchical of multiple hypergraphs, i.e. deep hypernetworks. As described in the previous subsection, the learning of deep hypernetworks contains three technical issues. For handling these issues, in this dissertation, we proposed three methods as follows:

i) the advanced model representation for visual-linguistic features

ii) the introduction of a hierarchy into hypernetworks for representing multiple levels of features and a Monte-Carlo simulation-based stochastic method for efficiently exploring a huge combinatorial feature space

iii) incremental learning based on sequential Bayesian update for tracing concept drifts.

A hypernetwork represents multimodal associations by denoting vertices and hyperedges to visual-textual features and higher-order connections between vertices using a flexible hypergraph structure, i.e., multimodal hypernetwork. By defining an image patch and a word as a visual and a textual node, a hyperedge encodes a semantic building block consisting of patches and words, and a multimodal hypernetwork represent multimodal association rules with a large population of many hyperedges to characterize the empirical distribution of observed multimodal data.

Since hypernetworks are represented by a huge combinatorial space, it is almost infeasible to explore the space with an exhaustive search considering high-dimensionality of multimodal data. For efficiently search the space of hypernetworks, we use a stochastic method for constructing hypergraphs based on the Monte-Carlo simulation, graph Monte-Carlo (graph MC). Graph MC constructs a hypergraph by probabilistically selecting vertices based on observed data to generate hyperedges. Also, learning from large-scale data generally causes the scalability problem because it requires huge number of hyperedges. As an alternative, we in-

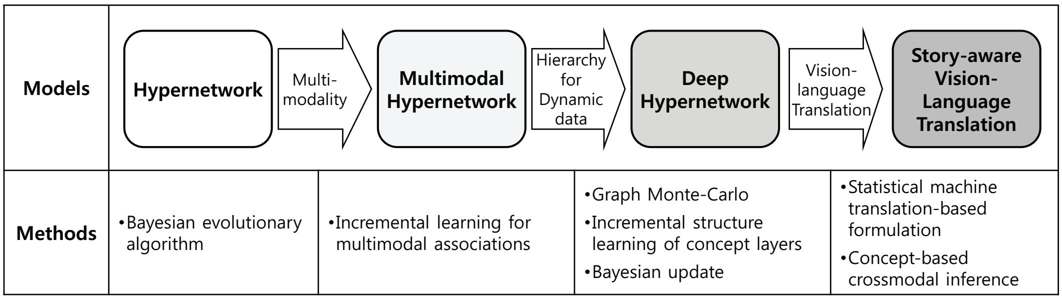

Models
Hypernetwork
Multi-modality
Multimodal Hypernetwork
Hierarchy for Dynamic data
Deep Hypernetwork
Vision-language Translation
Story-aware Vision-Language Translation
Methods
•Bayesian evolutionary algorithm
•Incremental learning for multimodal associations
•Graph Monte-Carlo
•Incremental structure learning of concept layers
•Bayesian update
•Statistical machine translation-based formulation
•Concept-based crossmodal inference

Figure 1.2: Improvement of the proposed models in this dissertation

troduce a deep architecture into the hypernetworks using a hierarchy of multiple hypergraphs, a deep hypernetwork, instead of the increase of the amount of hyperedges. Nodes in upper layers encode higher-level features, vice versa. Deep hypernetworks are different from other deep network models such as deep neural networks and deep Boltzmann machines in terms of the flexibility of the model structure. While nodes are fully connected between layers in conventional deep networks, the connectivity and the number of nodes of deep hypernetworks are sparse and flexibly change as the learning proceeds. This sparse and hierarchical structure reduces the model complexity, pursuing a parse modular hierarchical structure, as found in human brains (Quiroga, 2012).

Deep hypernetwork incrementally learns the knowledge by the graph MC and the weight update process while observing new data, thus robustly tracing concept drift and continuously accumulating new knowledge. This process is formalized as a sequential Bayesian inference. Finally, the learned model represents a hierarchy of high-level features, which can be considered as concept knowledge on the observed data. In this dissertation, using cartoon videos as the data, deep hypernetworks models concept hierarchies on the video contents and the constructed

Models and Methods	Data	Main Results
Multimodal Hypernetworks • Higher-order representation of visual-textual features • Incremental learning	**SBU Photograph DB** (Ordonez, 2011) - Images with descriptive sentences	• Image retrieval • Outperforming successful retrievals • **Retrieved images** on text queries
Deep Hypernetworks • Hierarchy of hypergraphs • Concept hierarchies • Construction of knowledge on video contents	**"Pororo"** - Cartoon video series - 14 DVDs and 183 episodes	• Representation of **visual-linguistic concepts** on the videos • Analyzing the **model structure change** and the **concept development**
Story-aware Vision-Language Translation • Machine translation-based formulation of V-L conversion • V-L crossmodal inference		• **Sentence-to-scene** generation • **Scene-to-sentence** generation • **Visual-linguistic video summarization**

Figure 1.3: Main results and used data for the proposed models

knowledge are used for tasks of converting between vision and language considering the contents. For achieving this, we formulate the vision-language conversion considering the contents in terms of the statistical machine translation, story-aware vision-language translation.

Figure 1.2 presents the improvement of the proposed models in this dissertation and Figure 1.3 summarizes the main experimental results.

1.4 Organization of the Dissertation

The rest of this dissertation is organized as follows:

- Chapter 2 presents a survey of the related work. Firstly, we discuss studies on multimodal learning. In particular, we summarize two representative methods for multimodal learning; non-parametric approach using topic models and deep learning-based models. Next, we explain hypernetworks as a higher-order model and the evolutionary learning method for hypernetworks in brief.

- In Chapter 3, we propose a multimodal hypernetwork which is extended to represent higher-order associations among visual and linguistic features and apply to text-to-image retrieval. The proposed multimodal hypernetwork. By defining image patches and words as vertices, multimodal hypernetworks characterize the association rules with the population of hyperedges encoding higher-order relationships between small images and words. The proposed model incrementally learns the associations from large-scale database of images with description. Given textual words, the query is transformed into visual words by visual-textual crossmodal query expansion. We evaluate the proposed method on SBU photograph data. The experimental results present that the proposed method achieves very competitive retrieval performances compared to a baseline method. Moreover, we demonstrate that our method provides robust text-to-image retrieval results for the increasing data.

- Chapter 4 shows a deep hypernetwork using a hierarchy of hypergraphs for representing concept hierarchies and handling concept drifts implicated in multimodal data. We propose the graph MC for efficiently learning deep hypernetworks. Using the graph MC, deep hypernetworks learns the concept hierarchies by flexibly changing the structure and the weights of hyperedges are incrementally updated to robustly trace concept drifts, which are formulated in terms of Bayesian inference. For evaluation, we use cartoon videos called "Pororo" as multimodal data and present the emergence and evolution of visual-linguistic concepts of the videos as the stories unfold. Also, we investigate how the deep architecture and the learning strategy of the graph

MC influence on the learning.

- In Chapter 5, we propose a method for story-aware vision-language translation based on the knowledge constructed by learning of deep hypernetworks. The story-aware translation transforms between scene images and subtitles considering the story contents learned and this process is formulated in terms of the statistical machine translation. We also evaluate the proposed method on ″Pororo″. Experimental results present that our method precisely translates the scenes into the subtitles and vice versa, reflecting observed video contents. In addition, investigate how the deep architecture and the learning strategy of the graph MC influence on the translation performance.

- This dissertation is summarized and directions for further research are discussed in Chapter 6.

Chapter 2

Related Work

2.1 Multimodal Leanring

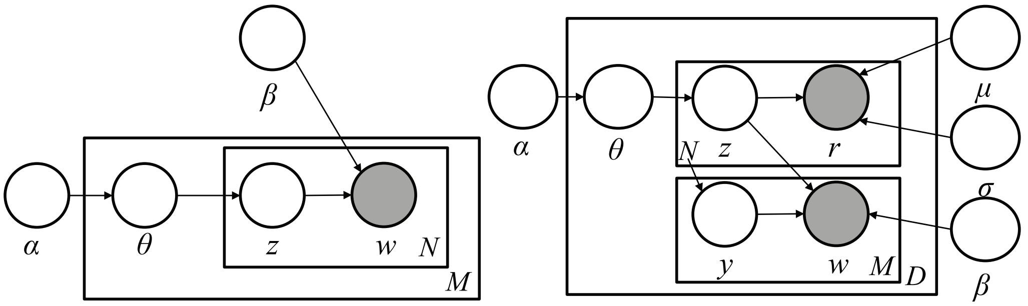
Multimodal learning is a method based on statistical machine learning for associating between two or more modal representation from multimodal data such as video, audio, and images. The research on multimodal associations have been studied for investigating human cognitive mechanisms, functional connectivity of brains, and psychological diseases in cognitive science (Mesulam, 1998; Zimmerman and Zeller, 1992; Meltzoff, 1990; Rioux and Van Meter, 1990) and neuroscience (Andersen et al., 1997; Halgren et al., 1994; Mesulam, 1994; Besson et al., 1990) since the early 1990s. These studies were mainly based on human experiments and made the great contributions to understand human cognition and brain function. Since the 2000s, many studies in computational cognitive and brain sciences mainly addressed the mulitmodal associations in terms of the Bayesian inference (Lee, 2011; Goodman et al., 2008; Kemp et al., 2006; Tenenbaum, 1999), related to the grounded theory of human cognition (Kiefer and Barsalou, 2012).

Apart from these scientific issues, the explosion of multimodal data enhance the importance of multimodal learning by which the information found can be used for more valuable applications and real-world problems in information retrieval (Carpineto and Romano, 2012), recommendation (Mei et al., 2011), and decision-making support (Fearon et al., 2012). Furthermore, multimodal learning is essential for learning and predicting human behaviors from log data generated by diverse sensors including body, emotion, and movement sensors (Gemmell et al., 2006; Eagle and Pentland, 2006), which is a key method for lifelong learning (Yu et al., 2014; Zhang, 2013) to achieve human-level artificial intelligence (Muggleton, 2014). Therefore, many conventional machine learning methods have been improved to efficiently deal with multimodal data such as Bayesian non-parametric methods (Blei, 2012), matrix factorization-based methods (Nikitidis et al., 2012; Caicedo et al., 2012), Markov networks (Karimaghaloo et al., 2012; Fan et al., 2011) and deep learning models (Kiros et al., 2014; Socher et al., 2013; Srivastava and Salakhutdinov, 2012), thus showing successful applications.

The rest of this chapter summarize state-of-the art models for multimodal learning and we discuss their characteristics and the limitations in brief.

2.2 Models for Learning from Multimodal Data

In this section, we summarize two approaches used widely for learning multimodal associations. One is approaches using topic models based on latent Dirichlet allocation (LDA) and the other is deep learning-based methods. Two models belong to probabilistic graphical models where associations between multimodal variables are defined as a probability and one representation is transformed into the other modality by inference processes.



(a) Latent Dirichlet allocation (b) Correspondence-Latent Dirichlet allocation

Figure 2.1: Graphical representation of the LDA model (a) and the corr-LDA model (b). Gray boxes denote observable variables.

2.2.1 Topic Model-Based Multimodal Leanring

A topic model is a statistical model for finding the topics, abstract variables, which occur in data and it was used for text mining from a collection of documents in early time. Although early topic models were proposed in the late 1990s (Hofmann, 1999; Papadimitriou et al., 1998), the most common methods are latent Dirichlet allociation (LDA) proposed by Blei et al. (Blei et al., 2003) and its extended models. In LDA, each document is viewed as a mixture of many topics, which are represented as latent variables. Each topic has probabilities of generating words in documents and two different topics have the different probability distributions of words from each other. A topic is not strongly defined semantically in many cases but used for the basis on the labels in supervised learning. Figure 2.1(a) shows the plate notation of the standard LDA model. In Figure 2.1, z and w are topic and word variables. α and β denote the parameters of the Dirichlet prior on the per-document topic distribution and the per-topic word distribution, respectively. θ is the topic distribution for a document.

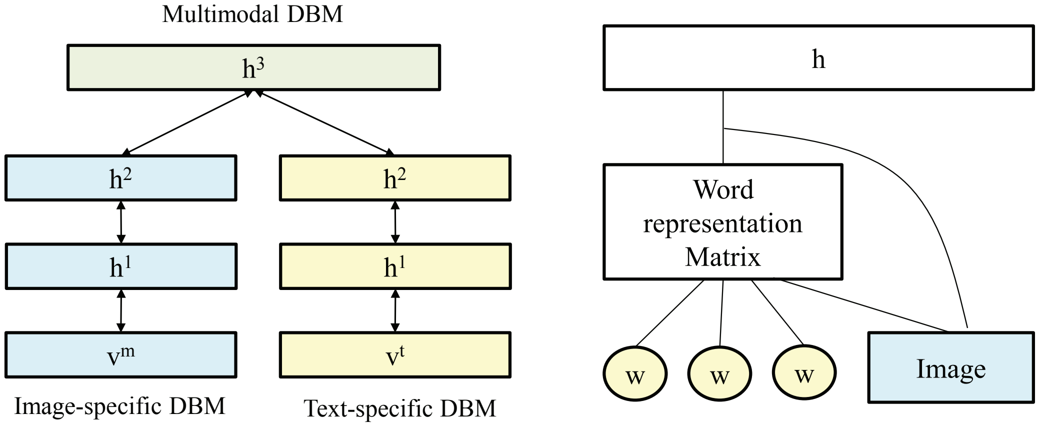
An LDA model was extended to model multimodal data such as annotated images by adding an additional plate including topics and observable variables into the model, i.e. corresponce-LDA (corr-LDA)(Blei and Jordan, 2003). Figure 2.1(b) presents the graphical representation of the corr-LDA model. In Figure 2.1(b), z and y denote the topics of images and words. In addition, r and w are small regions of an image regions and words. Each image region is represented as a real-valued vector of multiple visual properties and μ and σ are the mean and the standard deviation of the feature values. As shown in Figure 2.1(b), words are determined by the image topic as well as the word topic and the number of regions influences on the word topic in corr-LDA. These dependencies allow the model to associated with images and annotation words. LDAs were extended into an nonparametric model including hierarchical Dirichlet process mixture models (Teh et al., 2006) to applied mutimodal data recognition (Li and Fei-Fei, 2010; Guo and Wang, 2013), image annotation (Nguyen et al., 2013; Feng and Lapata, 2013), and cognitive modeling (Austerweil and Griffiths, 2013; Paddock and Savitsky, 2013).

Recent LDA-based topic models have been applied to video classification (Fu et al., 2014), object discovery in video frames (Zhao et al., 2013a), action recognition (Zhao et al., 2013b) and video pattern analysis (Jeong et al., 2014). These methods can be useful in real-world application such as video recommendation on the web and surveillance systems. However, topic models are difficult to intuitively understand what a topic means since the topics are not explictly identified. Therefore, it is not easy to extract and represent knowledge on the data contents from the learned models. In addition, it is not easy to apply incremental learning method because the models use a fixed model structure.

2.2.2 Deep Network-based Multimodal Leanring

Deep learning is one of the most successful methods in machine learning for the past ten years. It is algorithms which try to model high-level abstractions in data by introducing the hierarchical architecture composed of multiple non-linear transformation (Bengio et al., 2013). In addition, deep learning is a method based on learning representations of data to make it easier to learn tasks of interest, i.e., distributed representations.

Diverse deep learning method such as deep neural networks (LeCun et al., 1989), deep belief networks (Hinton, 2009; Le Roux and Bengio, 2008), deep Boltzmann machines (DBM) (Salakhutdinov and Hinton, 2012) and convolutional neural networks (Lawrence et al., 1997) have reported successful applications in computer vision, speech recognition, and natural language processing. A deep neural network (DNN) is an artificial neural network with multiple layers. General deep neural networks are typically designed as a feedforward network but recent studies on recurrent neural networks using deep learning architecture have reported successful applications as language models (Mikolov et al., 2010). Deep belief networks (DBN) are a generative probabilistic graphical model consisting of multiple layers of hidden units (Hinton, 2009). A DBN is efficiently trained in an unsupervised and layer-by-layer manner where the layer consists of restricted Boltzmann machines (RBM), which is an undirected and generative energy-based model with an input and single hidden layer. There is no connection between nodes of the same layer in an RBM. An RBM is trained based on contrastive divergence (CD) which provides an approximation to the maximum likelihood method. Once an RBM is tranied, another RBM is stacked to build a multilayer RBMs. Whenever the RBM is stacked, the input layer is initialized to a training vector and values for the units in the already-trained RBM layers. A convolutional neural network



(a) Multimodal deep Boltzmann machine (b) Multimodal neural language model

Figure 2.2: Graphical representation of the multimodal deep Boltzmann machine (a) and the multimodal neural language model (b)

(CNN) is a feed-forward network where the neurons are connected in such a way that they respond to overlapping regions in the visual field. CNNs consists of multiple layers of neuron collections corresponding to small portion of input images, called receptive fields. In particular, CNNs have been successfully applied in computer visions (Ji et al., 2013; Lauer et al., 2007; Phung and Bouzerdoum, 2007; Le Callet et al., 2006). Recent studies on deep learning models have been used for learning the associations between vision and language. Srivastava et al. proposed multimodal deep Boltzmann machines (MDBM) for associating images with texts (Srivastava and Salakhutdinov, 2012). MDBMs consist of two DBMs dedicated to images and texts and one DBM for joint representation of two modalities as shown in Figure 2.2(a). The mulimodal DBM was applied to image classification, text-to-image retrieval, and image annotation on a large-scale image databse (Huiskes and

Lew, 2008). Socher et al. proposed a zero-shot learning method method based on crossmodal transfer learning using deep network models and applied the model to image classification (Socher et al., 2013). Beyond image classification and annotation, the method for generating the sentences describing images was proposed by Kiros et al., called multimodal neural langauge model (Kiros et al., 2014) as shown in Figure 2.2(b). Ngiam et al. proposed a multimodal deep network for associating between video and audio and applied the model to video and audio reconstruction (Ngiam et al., 2011).

Despite many successful reports of deep learning models, their distributed representation makes the interpretation of the model difficult, thus being not suitable for representing knowledge. Also, they are not easy to apply incremental learning methods due to their fixed model structures.

2.3 Higher-Order Graphical Models

We generally assume pairwise relationships among the objects of our interest in machine learning problem setting. An object set endowed with pairwise relations can be naturally described as a graph, in which the vertices represent the objects, and two vertices related to each other are joined together by an edge. However, in many real-world problems, relationships among the objects are more higher-order than pairwise (Borgatti et al., 2009), and thus representing a set of their complex relationships as general undirected or directed graphs may not be complete. A higher-order model uses higher-order units as features . While linear models are difficult to reflect high order dependency embodied in the data, higher-order models can represent higher-order relationships, thus fitting the complex solution spaces including nonlinearity. A higher-order unit can be defined to a feature represented with patterns or function values derived from raw attributes of given data (Roddick et al., 2008; Lehár et al., 2008).

In this dissertation, we use an higher-order unit represented with a conjunction of attribute values of data, and a population of the units as a higher-order model. This conjunction-based unit representation enhances the interpretability of the models more compared with units based on numerical functions. Also, the individual and the population in our study can be represented with a hyperedge and a hypergraph.

2.3.1 Hypernetwork Models

Hypergraphs (Berge, 1984; Zhou et al., 2006)are a generalized graph in terms of the power of representation in graph theory. Whereas an edge in conventional graphs only represents a bi-relationship between two vertices, an edge in a hypergraph-a hyperedge-can connect two or more vertices concurrently. Formally, a hypergraph is defined as $G = (V, E)$, where V and E are a set of vertices v and a set of hyperedges e, respectively. A hyperedge is a subset of V and it has a weight, . Let $d(v)$ and $\delta(e)$ denote the degree of a vertex and a hyperedge, respectively. We define an indicator function $h(v, e)$ which returns 1 if v is an element of e and 0, otherwise. Then, the degrees, $d(v)$ and $\delta(e)$ are defined as

$$d(v) = \sum\nolimits_{e \in E} w(e)h(v, e) \quad and \quad \delta(e) = |e| \tag{2.1}$$

where $|e|$ is the cardinality (size) of e. A hyperedge of degree $\delta(e) = k$ is called a k-hyperedge. Higher-degree hyperedges characterize more specific patterns while lower-degree ones include more general information. When a hypergraph consists of k-hyperedges only, we call it a k-uniform hypergraph or k-hypergraph. Then, we can consider a conventional graph as a 2-hypergraph. Hypergraphs have been applied to modeling a variety of problems such as clustering (Zhou et al., 2006), text mining (Hu et al., 2008), multimedia mining (Tan et al., 2008), bioinformatics (Klamt et al., 2009), and building Markov logic networks (MLNs) (Kok and Domingos, 2009).

Hypernetworks (Zhang, 2008) are a higher-order model using a hypergraph structure. A vertex in a hypernetwork is defined as a pair of a data variable and its value, and a hyperedge corresponds to an arbitrary higher-order connection among vertices. The weight of a hyperedge reflects the strength of its connectivity. Since a hypernetwork is the population consisting of large number of hyperedges, therefore, the model characterizes higher-order relations among the variables. This hypernetwork representation provides the model with two advantages: flexible structure and interpretability. A hypernetwork has a very flexible structure compared to Genetic Programming (Koza, 1992) and neural trees (Zhang et al., 1997) because the degrees of hyperedges vary in the model and it is easy to add or remove new vertices and hyperedges into or from the model. In addition, connections among vertices allow the significant relationships to be easily extracted from the learned model by visualization. Since the hypernetwork was initially proposed as a simulation model for DNA molecular computing (Zhang and Jang, 2005a,b), they have been successfully used for diverse problems such as bioinformatics (Kim et al., 2013, 2010; Ha et al., 2007), pattern recognition (Kim and Zhang, 2007), disease prediction (Kim et al., 2014), multimodal information retrieval (Ha et al., 2010, 2009b), cortical data analysis, and cognitive modeling (Kim et al., 2011; Ha et al., 2009a) (Lee et al., 2013; Zhang et al., 2012).

A hypernetwork is a large population of many hyperedges including the class label and variable-value pairs. Formally, a hypernetwork is defined as a triple $H = (V, E, W)$, where W denotes a hyperedge weight set. A hyperedge in a hypernetwork is the set of two or more vertices including the class label:

$$e_i = \{v_{i1}, v_{i2}, v_{i3}, ..., v_{i|e_i|}, y_i\}, \tag{2.2}$$

where y_i is the class label of the i-th hyperedge e_i. This definition enables a hyperedge to be considered as a decision rule. The weights of hyperedges reflect their discriminative capability with respect to the class label. Thus, a hyperedge can be regarded as a weak learner characterizing the partial pattern necessary for classification, so a hypernetwork is an ensemble consisting of many weak learners. Fig. 1 shows the population of hyperedges and its corresponding hypergraph structure. When the n-th data instance $\mathbf{x}^{(n)}$, a class label set Y, and a hypernetwork H are given, the class label of the instance is then classified as whose weighted sum of hyperedges matched to $\mathbf{x}^{(n)}$ is largest among the elements of Y. Specifically, we determine the class label as follows:

1. Calculate the total weight $\tilde{w}_y$ as the summation of weights for $y \in Y$ with all hyperedges in the hyperedge set E such that

$$\tilde{w}_y = \sum_{i=1}^{|E|} \left\{ w(e_i) f(x^{(n)}, e_i) \varphi(y, y_i) \right\}, \tag{2.3}$$

where $w(e_i)$ denotes the weight of e_i.

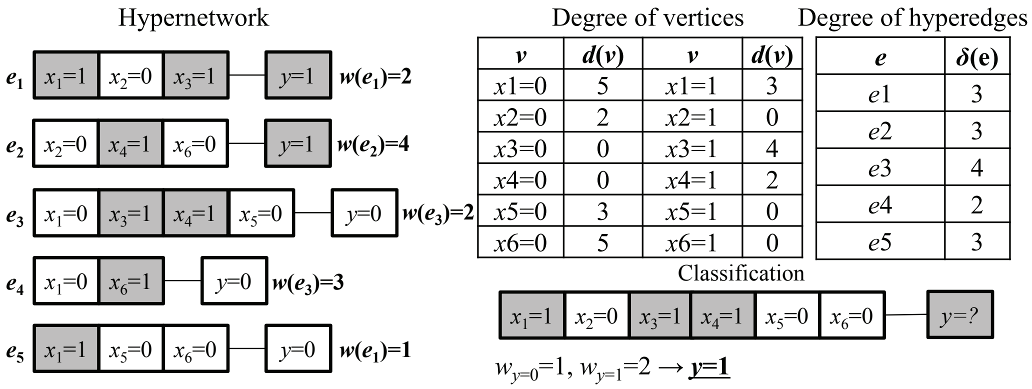
2. Predict $\hat{y}^{(n)}$ as the label of $\mathbf{x}^{(n)}$ that has the largest total weight:

$$\hat{y}^{(n)} = \arg\max_{y \in Y} \tilde{w}_y \tag{2.4}$$

$f(x^{(n)}, e_i)$ and $\varphi(y, y_i)$ are a matching function and an indicator function which return 1 if e_i matches $\mathbf{x}^{(n)}$ and if $y^{(n)} = y_i$, respectively as follows:

$$f_i^{(n)} = f(x^{(n)}, e_i) = \begin{cases} 1, \text{if} \exp\left\{ c(x^{(n)}, e_i) - \delta(e_i) \right\} > \theta \\ 0, \text{otherwise} \end{cases}, \tag{2.5}$$

$$\varphi_i^{(n)} = \varphi(y^{(n)}, y_i) = \begin{cases} 1, \text{if} y^{(n)} {=} y_i \\ 0, \text{otherwise} \end{cases}, \tag{2.6}$$


Hypernetwork
e_1 x_1=1 x_2=0 x_3=1 y=1 $w(e_1)$=2
e_2 x_2=0 x_4=1 x_6=0 y=1 $w(e_2)$=4
e_3 x_1=0 x_3=1 x_4=1 x_5=0 y=0 $w(e_3)$=2
e_4 x_1=0 x_6=1 y=0 $w(e_3)$=3
e_5 x_1=1 x_5=0 x_6=0 y=0 $w(e_1)$=1
Classification
x_1=1 x_2=0 x_3=1 x_4=1 x_5=0 x_6=0 y=?
$w_{y=0}$=1, $w_{y=1}$=2 → y=1

Degree of vertices

v	*d*(*v*)	*v*	*d*(*v*)
*x*1=0	5	*x*1=1	3
*x*2=0	2	*x*2=1	0
*x*3=0	0	*x*3=1	4
*x*4=0	0	*x*4=1	2
*x*5=0	3	*x*5=1	0
*x*6=0	5	*x*6=1	0

Degree of hyperedges

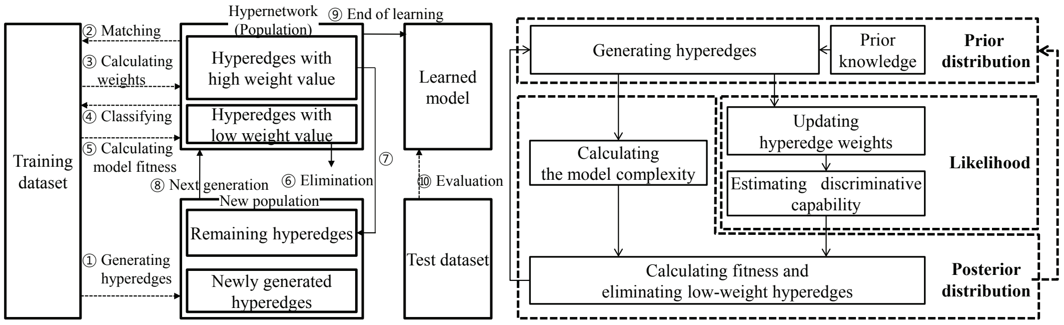
e	*δ*(e)
*e*1	3
*e*2	3
*e*3	4
*e*4	2
*e*5	3

Figure 2.3: An example of a hypernetwork and its term

where $c(x^{(n)}, e_i)$ denotes matching number, the number of hyperedge variables whose value is equal to the value of their corresponding variables in $\mathbf{x}^{(n)}$. θ is the matching threshold and plays the role of a smoothing factor, enhancing robustness against data noise by allowing partial matching. Figure 2.3 presents an instance of hypernetworks and their terms.

2.3.2 Bayesian Evolutionary Learning of Hypernetworks

Hypernetworks use an evolutionary algorithm for learning from data. As shown in Figure 2.4(a), the evolutionary algorithm for learning hypernetworks consists of generating hyperedges, updating the weight of hyperedges, and evaluating the model. When a training dataset is given, hyperedges are generated to construct an initial hypernetwork. The weight of hyperedges is calculated by matching the hyperedges with the training instances. After updating the weights, the fitness value of model is estimated by classifying the training set. Then, the model is evolved by replacing low weighted hyperedges with newly generated ones at every


Training dataset
② Matching
③ Calculating weights
④ Classifying
⑤ Calculating model fitness
① Generating hyperedges
Hypernetwork (Population)
Hyperedges with high weight value
Hyperedges with low weight value
⑧ Next generation
⑥ Elimination
New population
Remaining hyperedges
Newly generated hyperedges
⑨ End of learning
⑦
Learned model
⑩ Evaluation
Test dataset
Generating hyperedges
Prior knowledge
Prior distribution
Calculating the model complexity
Updating hyperedge weights
Estimating discriminative capability
Likelihood
Calculating fitness and eliminating low-weight hyperedges
Posterior distribution

(a) Evolving flow of hypernetworks

(b) Bayesian perspective of hypernetwork evolution

Figure 2.4: Overall procedure for learning Bayesian evolutionary hypernetworks. (a) presents the flow of learning hypernetworks with a functional level. (b) explains the evolution of hypernetworks from an Bayesian point of view.

generation. This evolutionary learning of hypernetworks can be defined as in terms of the sequential Bayesian inference and we call it Bayesian evolutionary computation.

Bayesian evolutionary computation (BEC) views evolutionary computation as a sequential Bayesian sampling process which transmits information from prior to posterior with likelihood estimation based on fitness measurements (Zhang, 2000). The evolved model in the posterior then plays the role of the empirical prior in the next generation. A hypernetwork is learned by BEC and is called a Bayesian evolutionary hypernetwork (BEHN) (Ha et al., 2014c; Kim et al., 2014). BEC assumes that the posterior and the prior are represented as the current and the previous populations. Specifically, the model fitness is defined as the posterior probability which reflects both data-discrimination capability and the model complexity. This definition of the fitness allows the model to efficiently search the huge space and to adaptively determine the model complexity in BEHNs.

Let H_t be a BEHN at the t-th generation. For a dataset D, the posterior distribution of H_t is given by Bayes' rule as follows:

$$p(H_t|D) = \frac{p(D|H_t)p(H_t)}{p(D)} \tag{2.7}$$

For a classification problem, the dataset is decomposed as $D = (X, Y)$, where $X = \{x^{(n)}\}_{n=1}^{N}$ and $Y = \{y^{(n)}\}_{n=1}^{N}$. Then, the classification rule is given as the conditional probability:

$$p(H_t|X, Y) = \frac{p(Y|X, H_t)p(H_t|X)}{p(Y|X)} \tag{2.8}$$

where $p(Y|X, H_t)$ and $p(H_t|X)$ are called the likelihood and the prior, respectively. Also, $p(Y|X)$ is a normalizing constant since it is not a function of H_t. Thus, the posterior distribution is proportional to the product of the likelihood and the prior:

$$p(H_t|X, Y) \propto p(Y|X, H_t)p(H_t|X) \ . \tag{2.9}$$

We define the fitness F_t of H_t as the logarithm of the posterior so that the evolutionary process is to maximize it:

$$F_t = \log p(Y|X, H_t) + \log p(H_t|X) \ \ and \ \ H* = \arg\max_{H_t} F_t \tag{2.10}$$

Figure 2.4(b) illustrates the evolving process of BEHNs with the perspective of BEC. The likelihood is defined as the conditional probability of correctly classifying Y from a model H and X and is considered as the discriminative capability. For estimating the likelihood, we assume that the discriminative capability grows by increasing the difference of the weighted summation between the correctly matched hyperedges and the incorrectly matched hyperedges for all training data. Also, the prior is defined to prefer less complex model structure, which means a model consisting of smaller number of hyperedges. The iteration of sequential Bayesian evolutionary process thus finds optimal composition and number of the hyperedges, that is, a hypernetwork that increases the classification accuracy while keeping the model complexity as sparse as possible. Although BEHNs represent higher-order feature relations and efficiently learn from high-dimensional data, they have focused on classification problems using supervised learning as a discriminative model rather than a generative model.

Chapter 3

Multimodal Hypernetworks for Text-to-Image Retrievals

3.1 Overview

Text-to-image (T2I) retrieval (Datta et al., 2008) involves getting images from text queries and it has been actively studied because of its diverse applications including content-based image retrieval (Datta et al., 2008; Smeulders et al., 2000) and article or video searching (Feng et al., 2004). Various approaches have been applied to associate textual and visual modalities for T2I retrieval. Feng et al. used multiple Bernoulli model for image retrieval (Feng et al., 2004) and Zhang et al. applied Bayesian framework to learning latent semantic models for T2I retrieval (Zhang et al., 2005). Li et al. proposed multi-instance learning method using loosely labeled images for image retrieval (Li et al., 2011). Because the sizes of text and visual vocabularies increase continuously due to the growth of multimodal data, however, T2I retrieval models should facilitate to deal with these increasing data for their practical usages, thus requiring an incremental learning method. However, most of the models for T2I retrieval assume the vocabulary sizes are fixed like bag of words and they use batch approach-based learning methods (Feng et al., 2004; Zhang et al., 2005; Li et al., 2011). Therefore, these models are not easy to be practically applied to data-increasing environments since this fixed vocabulary size has a limitation in representing new multimodal image data including unobserved textual words and new visual features.

In this chapter, we propose a T2I retrieval method based on a textual-visual association model which can efficiently treat increasing data (Ha et al., 2012). For the multimodal association, we use a hypernetwork (HN), which is a higher-order probabilistic graphical model using hypergraph structure (Zhang, 2008). In HNs, a vertex denotes a textual word or a visual feature and a hyperedge represents a multimodal subpattern of textual-visual data by connecting more than two vertices. Therefore, HNs can represent the higher-order associative relationships among textual and visual modalities. Learning HNs consists of generating hyperedges which reflect the relationships embodied in the given data and updating the weights of the hyperedges. This learning process is formulated by a sequential Bayesian framework. Whenever an image with a description is observed, new hyperedges are generated from the image by random selection-based evolutionary method and they are added into the HN. Then, the weights of the hyperedges of the model are updated by predicting and correcting the observed image, with comparing the subpattern of each hyperedge with the image and its description. Therefore, the weights are to reflect the associative strength of the hyperedge for predicting the images and descriptions. Especially, HNs can incrementally learn the increasing data by simply adding unobserved textual words and visual features involved in new data as new vertices into the model and generating hyperedges including them. When a text query is given, the query is expanded to a visual query consisting of

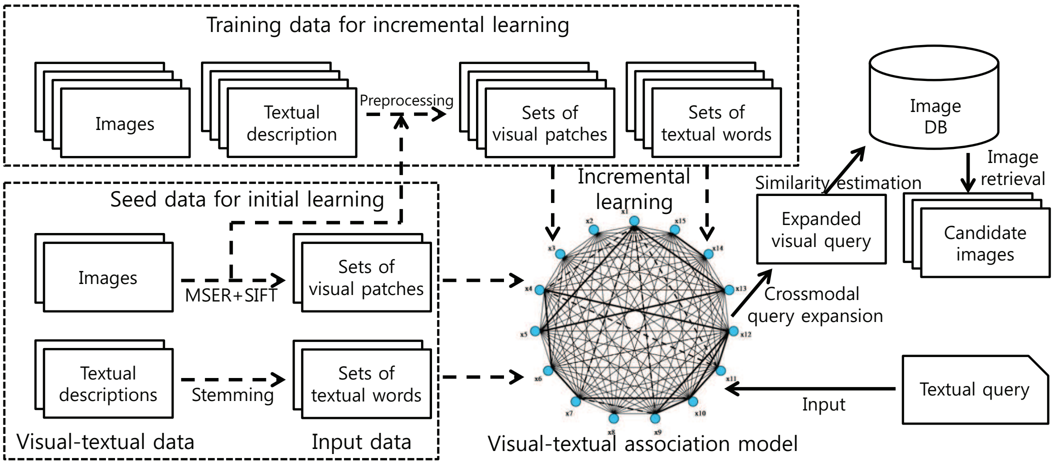


Figure 3.1: Overall flow of text-to-image retrieval using incrementally learned multimodal hypernetworks

visual patches associated with the textual query by the learned HNs. By measuring the similarity between the expanded visual query and stored images, images are retrieved semantically related to the given text query. Figure 3.1 describes the proposed framework of T2I retrieval using HN models.

We apply the HN-based T2I retrieval method to retrieve about 3,000 images from Flickr.com for evaluation. In this study, several visual patches are extracted from an image by maximally stable external regions (MSER) (Matas et al., 2004) and the extracted patches are represented with 500 scale-invariant feature transform (SIFT) features (Lowe, 2004). Also, the image descriptions are represented with about 2,800 textual words. The experimental results present that our method shows good retrieval performances over a baseline method based on the co-occurrence of textual words and visual features. Moreover, we demonstrate that the proposed method provides robust T2I retrieval results with reflecting the increase of the data.

3.2 Hypernetworks for Multimodal Associations

3.2.1 Multimodal Hypernetworks

The previous works proposed a method for multimodal associations between texts and images (Ha et al., 2009b, 2010). However, they contains the semantic gap problem because the models use gray-scale pixels and SURF histrogram vectors as visual features, respectively. For solving this problem, we use small image patches as more semantic visual features. Then, an HN can be used as a multimodal association model by defining vertices to textual words or image pathces and hyperedges to associative relationships among textual and visual features (Zhang et al., 2012). The advantages of HNs as a multimodal association model are summarized as follows: i) Representation of multimodal association based on higher-order relationships among textual and visual features, ii) Robust and flexible model structure suitable for incremental learning, iii) Crossmodal inference based on higher-order associative strength for text-to-image retrieval. Figure 3.2 illustrates hyperedges generated from a captioned image. As shown in Figure 3.2, each hyperedge represents a higher-order visual-textual association by consisting of the several visual and textual subpatterns. Moreover, the HN has the flexible model structure for incremental learning because new textual words and visual patches involved in unobserved data are added as new vertices and the relationships between the new vertices are included as new hyperedges into the model. When a captioned image dataset $D = \{(\mathbf{x_T}, \mathbf{x_I})^{(n)}\}_{n=1}^{N}$, where $\mathbf{x_T}$ denotes the set of textual words in an image description and $\mathbf{x_I}$ is the set of visual patches comprising the image, is sequentially given, an HN can be formally considered as a mixture model of many hyperedges and the empirical distribution is represented with the model:

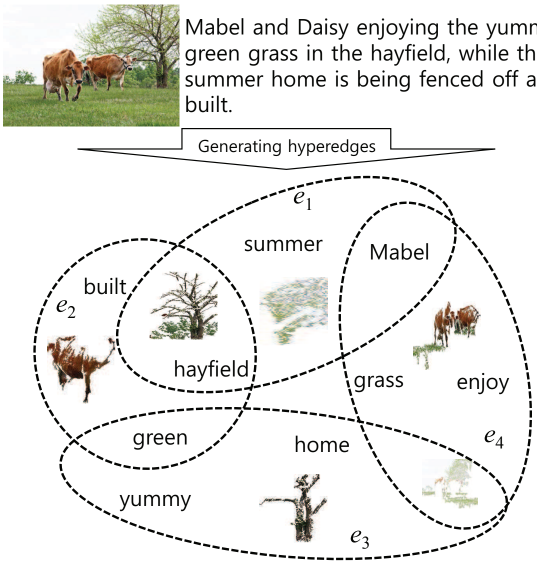

Mabel and Daisy enjoying the yummy green grass in the hayfield, while their summer home is being fenced off and built.
Generating hyperedges
e_1
summer
Mabel
built
e_2
hayfield
grass
enjoy
green
home
e_4
yummy
e_3

Figure 3.2: Hyperedges consisting of three texutual words and two visual pateches from a textual-visual data instance (a captioned image)

$$p(\mathbf{x_T}, \mathbf{x_I}|H) = \sum_{e \in E} w(e) f(\mathbf{x_T}, \mathbf{x_I}|e) \tag{3.1}$$

s.t.

$$0 < w(e) < 1 \quad \text{and} \quad \sum_{e \in E} w(e) = 1$$

,

where H denotes an HN model, $w(e)$ denotes the weight of a hyperedge e, and $f(\mathbf{x_T}, \mathbf{x_I}|e)$ is the density function. Then, the likelihood of the model is the probability of regenerating the observed data from the model and it is defined to this empirical distribution:

$$
\begin{aligned}
p(D|H) &\approx p(\mathbf{x_T}^{(1)}, \mathbf{x_I}^{(1)}, \mathbf{x_T}^{(2)}, \mathbf{x_I}^{(2)}, ..., \mathbf{x_T}^{(N)}, \mathbf{x_I}^{(N)}|H) \\
&= \prod_{n=1}^{N} p((\mathbf{x_T}, \mathbf{x_I})^{(n)}|H) = \prod_{n=1}^{N} \sum_{e \in E} w(e) f(\mathbf{x_T}, \mathbf{x_I}|e) \qquad (3.2)
\end{aligned}
$$

where $\mathbf{x_T}^{(n)}$ and $\mathbf{x_I}^{(n)}$ denote the textual description and the image of the n-th captioned image, respectively.

3.2.2 Incremental Learning of Multimodal Hypernetworks

Whenever observing an unseen descripted image ($\mathbf{x_T}$,$\mathbf{x_I}$), an HN is learned incrementally by predicting the image and updating the weight of the hyperedges. This learning procedure can be formulated by Bayes rule:

$$
p(H_n|\mathbf{x_T}, \mathbf{x_I}) = \frac{p(\mathbf{x_T}, \mathbf{x_I}|H_n) p(H_n)}{p(\mathbf{x_T}, \mathbf{x_I})} \qquad (3.3)
$$

where H_n denotes the HN at time step n. By this rule, the prior $P(H_n)$ is updated to the posterior $P(H_n|\mathbf{x_T}, \mathbf{x_I})$ by estimating the likelihood $P(\mathbf{x_T}, \mathbf{x_I}|H_n)$ and by normalized with $P(\mathbf{x_T}, \mathbf{x_I})$. The posterior is then used as the prior $P(H_n + 1)$ at the next time step n+1. Reformulating this process recursively using all time steps on the sequence of n data, the above equation is described:

$$
p(H_n|\mathbf{x_T}^{(1:n)}, \mathbf{x_I}^{(1:n)}) = \frac{p(\mathbf{x_T}^{(n)}, \mathbf{x_I}^{(n)}|H_n) p(H_{n-1}|\mathbf{x_T}^{(1:n-1)}, \mathbf{x_I}^{(1:n-1)})}{P(\mathbf{x_T}^{(n)}, \mathbf{x_I}^{(n)}|\mathbf{x_T}^{(1:n-1)}, \mathbf{x_I}^{(1:n-1)})} \qquad (3.4)
$$

$$
p(H_n|\mathbf{x_T}^{(1:n)}, \mathbf{x_I}^{(1:n)}) \propto p(\mathbf{x_T}^{(n)}, \mathbf{x_I}^{(n)}|H_n) p(H_{n-1}|\mathbf{x_T}^{(1:n-1)}, \mathbf{x_I}^{(1:n-1)}) \qquad (3.5)
$$

where $\mathbf{x_T}^{(1:n)}$ and $\mathbf{x_I}^{(1:n)}$ denote the sequential stream of n textual-visual data. The posterior is estimated by predicting the new image with both hyperedges generated from the new observed image and hyperedges of $H_n - 1$ learning *n-1* images. Each hyperedge is generated by randomly selecting visual patches of the given image and textual words of the description. This generation method assures that there always exists the subpattern involved in the hyperedge in the data. The details of generating hyperedges are explained in (Zhang et al., 2012) and (Ha et al., 2010). The weights of the hyperedges are updated by the prediction of the new observed image:

$$w_n(e) = \eta g(e, (\mathbf{x_T}, \mathbf{x_I})^{(n)}) + (1-\eta) w_{n-1}(e) \tag{3.6}$$

$$= \eta g_T(e_T, \mathbf{x_T}^{(n)}) \cdot g_I(e_I, \mathbf{x_I}^{(n)}) + (1-\eta) w_{n-1}(e) \tag{3.7}$$

s.t.

$$g_T(e_T, \mathbf{x_T}) = |e_T \cap \mathbf{x_T}|$$

and

$$g_I(e_I, \mathbf{x_I}) = \alpha \sum_{\mathbf{u} \in e_I} \max_{\mathbf{v} \in \mathbf{x_I}} \frac{A(\mathbf{u})A(\mathbf{v}^{\mathrm{T}})}{||A(\mathbf{u})||_0^1} + (1-\alpha) \sum_{\mathbf{u} \in e_I} \sum_{\mathbf{v} \in x_I} c(\mathbf{u}, \mathbf{v})$$

where is the constant for the current image, and e_T and e_I denote the sets of textual words and visual patches included in hyperedge e, respectively. Also, $\mathbf{u}$ and $\mathbf{v}$ denote the visual patches of e_I and $\mathbf{x}_I$, respectively, $A(\mathbf{u})$ is the function which returns the occurrence vector of SIFT features of $\mathbf{u}$, and denotes *L0*-norm of $A(\mathbf{u})$, the number of non-zero variables of $A(\mathbf{u})$. In addition, we add $c(\mathbf{u}, \mathbf{v})$ for reflecting the color similarity between two patches because SIFT does not consider a color property.

3.3 Text-to-Image Crossmodal Inference

An HN facilitates to translate text to image and vice versa by crossmodal inference because the model is the population of textual-visual associative subpatterns. In this paper, we focus on text-to-image translation for image retrieval from textual queries.

3.3.1 Representatation of Textual-Visual Data

The description sentences are represented to the subsets of textual word set used in the training image descriptions by stemming and eliminating the stop words. An image is represented to the set of several visual patches that are extracted by combining two image processing methods: maximally stable external regions (MSER) (Matas et al., 2004) and scale-invariant feature transform (SIFT) (Lowe, 2004). MSER is a method for detecting an invariant stable subset of external regions of the images and SIFT is a method for extracting the distinctive invariant features from the images. The given images are separately represented to the several external regions by MSER and the set of salient features by SIFT. The regions are then represented with the vectors of the SIFT features using their locality information and we use the SIFT-based regions as the visual patches. For effectively representing the visual patches with SIFT features, *k*-means clustering method is used in this study because there are few features shared by the regions when images are represented with raw SIFT features. For incremental learning, in addition, the visual patches of a new image are represented with the clustered SIFT features of the observed images by using the clustered features as the centroids for *k*-means method and by clustering the raw features of the new image with the centroids. Figure 3.3 illustrates the flow of converting an image into a visual patch set.

3.3.2 Text-to-Image Query Expansion

Text-to-image retrieval formally involves calculating the retrieved probability of an image $\mathbf{x_I}$ when a learned model H and a textual query $\mathbf{x_T}$ are given using (3.2):

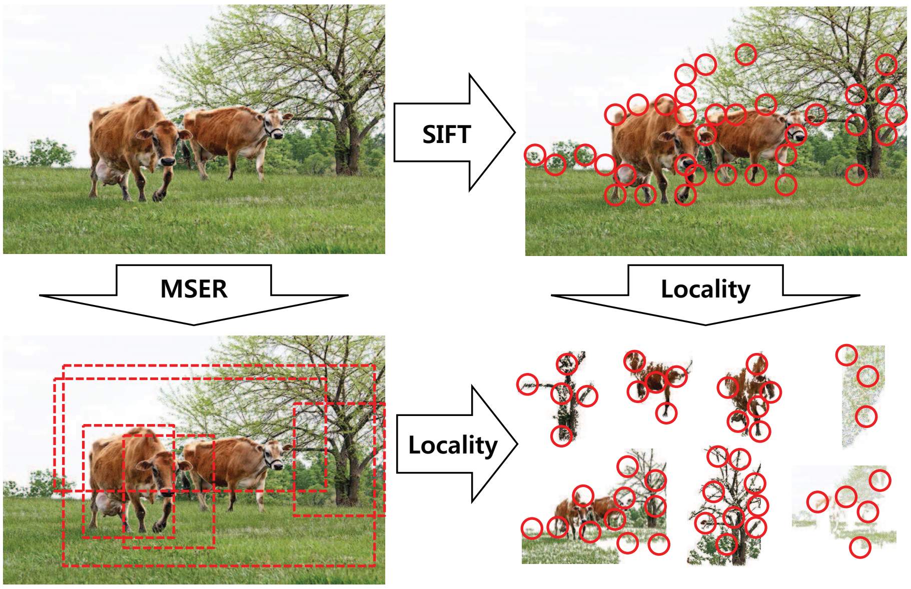


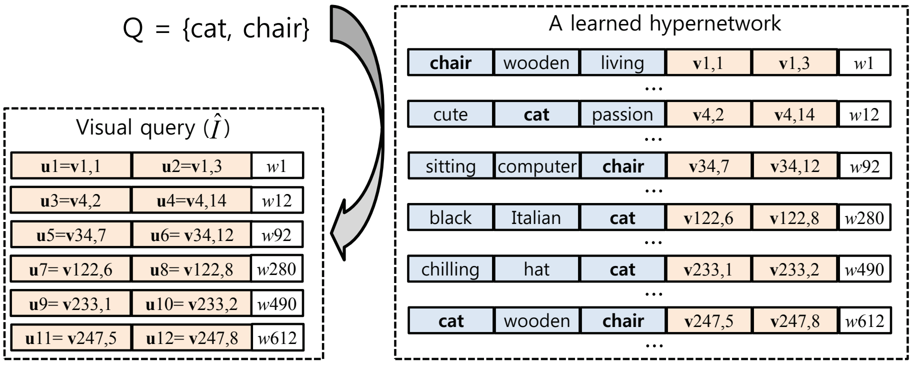
Figure 3.3: Flow of constructing visual patches from a given image with MSER and SIFT, two feature extraction methods. Boxes and circles denote the external regions extracted by MSER and the salient features by SIFT, respectively

$$p(\mathbf{x_I}|\mathbf{x_T}, H) = \frac{p(\mathbf{x_T}, \mathbf{x_I}|H)}{p(\mathbf{x_T}|H)} \propto \sum_{e \in E} w(e) f(\mathbf{x_T}, \mathbf{x_I}|e) \tag{3.8}$$

Then, the images related to the given textual words are selected as follows:

$$I^* = \arg\max_{\mathbf{x_I}} p(\mathbf{x_I}|\mathbf{x_T}, H) = \arg\max_{\mathbf{x_I}} \sum_{e \in E} w(e) f(\mathbf{x_T}, \mathbf{x_I}|e) \ . \tag{3.9}$$

When textual words are given as a query, in order to find I^*, we use the textual-to-visual query expansion method and several images are selected as the candidates of I* which are most similar to the visual query crossmodally expanded from the



u1=v1,1	u2=v1,3	w1
u3=v4,2	u4=v4,14	w12
u5=v34,7	u6= v34,12	w92
u7= v122,6	u8= v122,8	w280
u9= v233,1	u10= v233,2	w490
u11= v247,5	u12= v247,8	w612

chair	wooden	living	v1,1	v1,3	w1
…					
cute	cat	passion	v4,2	v4,14	w12
…					
sitting	computer	chair	v34,7	v34,12	w92
…					
black	Italian	cat	v122,6	v122,8	w280
…					
chilling	hat	cat	v233,1	v233,2	w490
…					
cat	wooden	chair	v247,5	v247,8	w612
…					

Figure 3.4: Flow of the crossmodal query expansion from the given textual query to the visual query

given textual query. Thus, (3.9) is reformulated by substituting $\mathbf{x_T}$ for the textual query Q as follows:

$$I^* = \arg\max_{x_I} \sum_{e \in E} w(e) f(Q, x_I | e) \approx \arg\max_{x_I} \delta(\hat{I}, x_I) \tag{3.10}$$

where $\hat{I}$ denotes the visual query expanded from Q and $\delta(\hat{I}, x_I)$ denotes a similarity function. Formally, a visual query $\hat{I}$ is defined to the set of visual patches involved in hyperedges including the elements of the textual query $Q = \left\{q_1, \ldots, q_{|Q|}\right\}$:

$$\hat{I} = \bigcup_{e \in E} \{u | u \in e, e \in E, Q \cap e \neq \emptyset\} \tag{3.11}$$

where $\mathbf{u}$ denotes visual patches. Figure 3.4 illustrates an example of the crossmodal query expansion from the textual query to the visual query. Then, the candidate images are retrieved by measuring the similarity between the expanded visual query and the stored images. The similarity is estimated by summing the similarity among the patches involved in $\hat{I}$ and $\mathbf{x_I}$:

$$\delta(\hat{I}, \mathbf{x_I}) = \frac{1}{|x_I|} \sum_{v \in x_I} \sum_{u \in \hat{I}} w(u) s(u, v) \tag{3.12}$$

s.t.

$$s(u, v) = \begin{cases} g_I(u, v), & if\, g_I(u, v) > \theta \\ 0, & \text{otherwise} \end{cases}$$

and

$$g_I(u, v) = \alpha \frac{A(u)A(v^{\mathrm{T}})}{\|A(u)\|_0^1} + (1 - \alpha)\, c(u, v)$$

where w(u) is the weight of the hyperedge including a visual patch u, $|\mathbf{x_I}|$ is the number of patches in $\mathbf{x_I}$, and denotes the threshold to prevent many low-valued patches from distorting the similarity. Therefore, the similarity becomes larger when the image involves the visual patches sharing more SIFT features with the patches of the visual query. Then, the images with large are retrieved as candidate images related to the textual query.

3.4 Text-to-Image Retrieval via Multimodal Hypernetworks

3.4.1 Data and Experimental Settings

We evaluate the proposed T2I retrieval method with the dataset consisting of 3,000 photography images described by the sentences from Flickr.com. For evaluation, we divide the dataset into training set and test set consisting of 1,000 and 2,000 images, respectively. Each description is represented with the subset of 2,814 textual words. An image is converted into the set of the visual patches represented with the occurrence vector of 500 clustered SIFT features. Table I shows parameter setups for the method.

Table 3.1: Parameter setup for model learning

Parameters	values
Number of visual patches in a hyperedge	2
Number of textual words in a hyperedge	3
Number of hyperedges generated from an image	5
Number of iterations for correction	5
α (constant for balancing SIFT and color)	0.99
θ (patch similarity threshold)	0.9

3.4.2 Text-to-Image Retrieval Performance

We use three measures such as precision, recall, and successful retrieval (SR) for evaluating the performance of the HN-based T2I retrieval method. In order to define the measures, we call it correct retrieval (CR) that a retrieved image explicitly includes the object that is described by the given textual query, i.e., query-object. Then, each measure is defined as follows:

$$precision = \frac{\#\,\mathrm{of\,CR}}{\#\,\mathrm{of\,the\,retrieved\,images}} \tag{3.13}$$

$$recall = \frac{\#\,\mathrm{of\,CR}}{\#\mathrm{of\,all\,images\,including\,the\,query} - \mathrm{object}} \tag{3.14}$$

$$SR = \begin{cases} 1, & precision > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

Moreover, we use two types of textual queries such as 10 queries and 30 queries for measuring the performance, and the textual queries are enumerated in Table

Table 3.2: Contents of two types of textual queries

Query	Textual words
10 queries	beach, boat, cat, flower, girl, grass, sand, sky, tree, water
30 queries	dog, dress, fish, floor, flower, girl, grass, house, kitchen, mountain, office, river, road, rock, room, sand, sky, table, tower, tree, wall, water

Table 3.3: Precision and recall of text-to-image retrieval for 10 queries

Methods	Precision	Recall
HN-based (1000-HN)	0.24	0.055
Baseline	0.155	0.035

3.2. Table 3.3 presents the precision and recall of T2I retrieval of the HN-based method with models learning all the training images (1000-HNs) for 10 queries on the test set compared to a baseline method. The baseline method uses all visual patches of the training images including the textual query in their description as the visual query without any learning process. From Table 3.3, the proposed T2I method outperforms the baseline method in terms of both precision and recall. The performances of the baseline method are lower than those of the proposed method since the expanded query of the baseline method is blurred by too many patches and the specificity is thus weakened. Meanwhile, large-weighted hyperedges with the strong associative relationships only survive in the HN by the incremental learning. Figure 3.5 presents the average precision and recall of T2I retrieval on the training set and the test set for 30 textual queries with 1000-HNs. As shown in Figure 3.5,

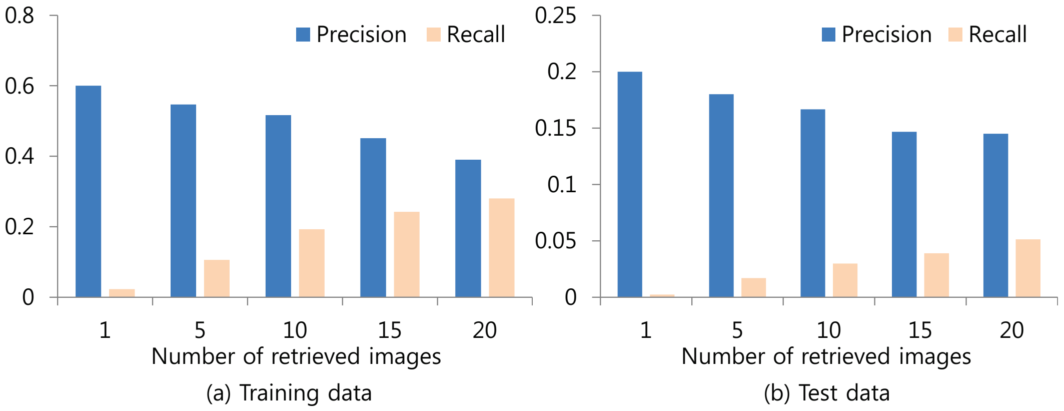


Figure 3.5: Precision and recall of text-to-image retrieval using 1000-hyperentworks for 30 queries on (a) the training dataset and (b) the test set. Values are averaged for 30 queries

Table 3.4: Successful retrieval for 30 queries

Image size	1	5	10	15	20
Training	0.6	0.933	1.0	1.0	1.0
Test	0.2	0.567	0.733	0.8	0.9

the results show the general pattern that the precision slightly decreases and the recall increases as the number of retrieved images grows up. For the training set, the precisions are mostly larger than 0.5 and it means that more than half of the retrieved images are related to the textual queries. From Figure 3.5(b), we indicate that one or more images are associated with the given query when the number of the retrieved images is 10 from the test set. Table 3.4 shows the SR of T2I retrieval for the same queries and model as Figure 3.5. Therefore, our method can retrieve

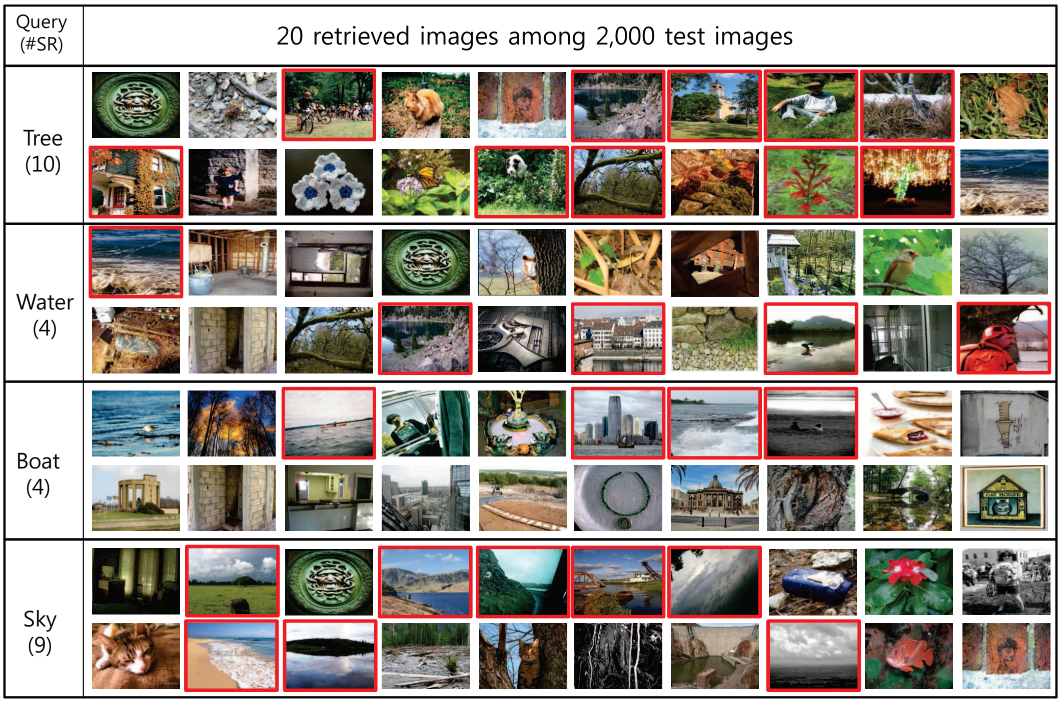


Figure 3.6: Retrieved images from the test dataset using 1000-hypernetworks for three queries. Red boxes are images including query-objects

images related to the textual query even if the images have no textual description. Figure 3.6 illustrates 20 retrieved images from the test dataset for each query, totally 60 images for three queries including 'tree', 'sky', and 'boat'. The precisions of 'sky' and 'tree' are higher than that of 'boat' because the patches of 'sky' and 'tree' are similar to each other due to sharing more SIFT features than the patches of boat.

Figure 3.7 shows the precision and the recall for 10 queries on the test set as the retrieval size increases. Although the performances are better than those of 30 queries because textual words in 10 queries are more selective, the result in Figure 3.7 is also consistent to Figure 3.5 In terms of SR, therefore, we indicate that images

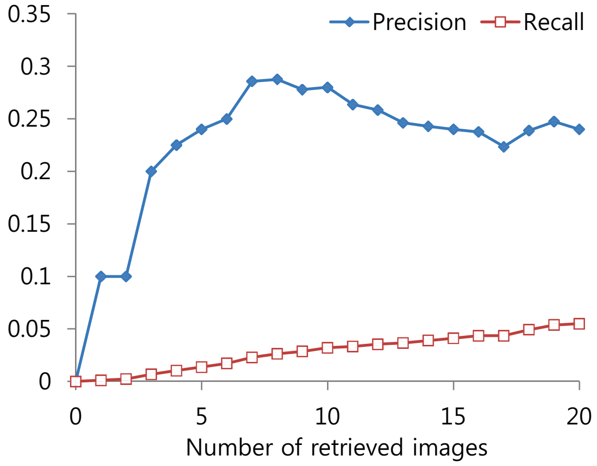


Figure 3.7: Precision and recall of text-to-image retrieval using 1000-hypernetworks for 10 queries on the test set as the retrieval size grows

associated with the text query can be retrieved with textual-visual crossmodal association by the HN-based T2I retrieval method from Table 3.4 and Figures 3.5, 3.6, and 3.5.

3.4.3 Incremental Learning for Text-to-Image Retrieval

Figure 3.8 shows (a) the sizes of the textual and the visual vocabularies of the model and (b) the performances of T2I retrieval, as the learning incrementally proceeds. As the observed data grow up, from Figure 3.8(a), visual information increases linearly due to the uniqueness of the patch while the textual vocabulary grows up slowly due to the frequently used words. In terms of performance, the precision and the recall are enhanced in early learning steps due to the increase of the number of hyperedges as well as the textual and the visual vocabulary sizes. Meanwhile,

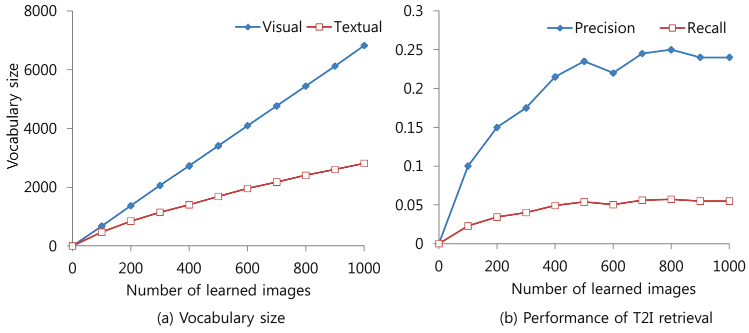


Figure 3.8: (a) Vocabulary size of hyperentworks and (b) average performance of text-to-image retrieval for 10 queries as learning proceeds incrementally. In (b), the number of retrieved images is 20.

the performances are saturated after learning more than 500 training images. The reason is that the model contains redundant information despite the uniqueness of the visual patches because the patches including the same object share many SIFT features. In addition, we can indicate that the decay of the performances is caused by the patches sharing SIFT features but including the different objects and this issue can be solved by specifically representing the visual patches with more SIFT features. Figure 3.9 illustrates the retrieved images for two textual queries including ’sky’ and ’grass’ as the increase of the observed images. Same as Figure 3.8(b), the model observing more images shows the higher retrieval performance for both queries in Figure 3.9.

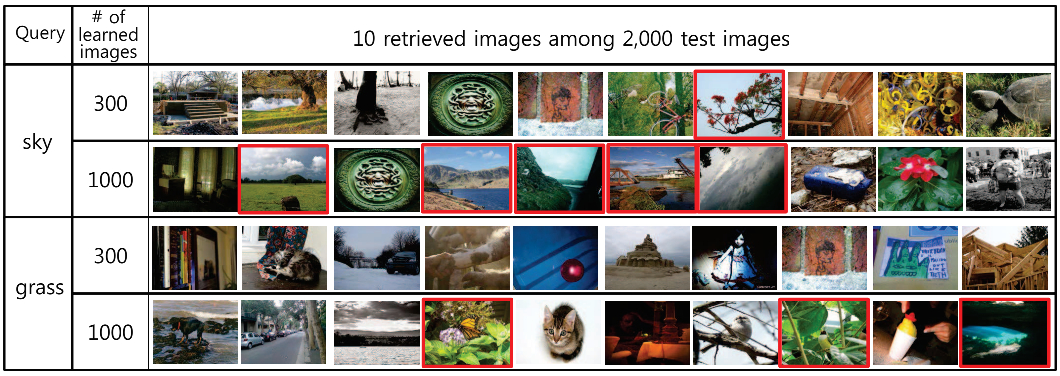


Figure 3.9: Retrieved images on the test dataset for two textual queries such as 'sky' and 'grass' with 300-hypernetworks and 1000-hypernetworks. Red boxed photos are images including the query-object.

3.5 Summary

We have proposed a novel text-to-image (T2I) retrieval method based on a textual-visual association model and we use hypernetwork models for incrementally learning the associative relationships between two modalities. Moreover, the images related to the text queries are retrieved by crossmodal query expansion with the learned model. The proposed method was evaluated on 3,000 images with text descriptions from Flick.com to retrieve images associated with various text queries. Experimental results show that our method achieves the high retrieval performance in terms of precision, recall, and successful retrieval on the test dataset compared to a baseline method. Moreover, the results demonstrate that the hypernetworks can learn robustly increasing data with the proposed incremental learning method.

Chapter 4

Deep Hypernetworks for Multimodal Cocnept Learning from Cartoon Videos

4.1 Overview

Recent explosion of data enhances the importance of automatic knowledge acquisition and representation from big data. Linguistically-oriented representation formalisms such as semantic networks (Steyvers and Tenenbaum, 2005) and WordNet (Fellbaum, 1998) are popular and extremely useful. However, mutually-grounded vision-language concepts are more foundational for cognitive systems that work in perception-action cycles. Existing text-oriented representations are inefficient for learning multimodal concepts from large-scale data, such as videos. Continuous knowledge construction from multimodal data streams is essential for achieving human-level artificial intelligence based on lifelong learning (Muggleton, 2014; Zhang, 2013).

The task of vision-language learning is to automatically build the relationships between vision and language from multimodal sources of data. Previous works on multimodal learning have focused on either cognitive theory or practical applications. On the practical side, the latent Dirichlet allocation (LDA) models were applied to image annotation (Blei and Jordan, 2003) and video object detection (Zhao et al., 2013a). Recently, deep learning models were also used for image annotation (Srivastava and Salakhutdinov, 2012) and descriptive sentence generation (Kiros et al., 2014). However, they mainly focused on automatic annotation rather than constructing semantic knowledge at a higher level. Furthermore, the techniques mostly have concentrated on efficient learning from a static large-scale dataset (Ordonez et al., 2011; Deng et al., 2009) but seldom considered the dynamic change of the contents, i.e. concept drift. Some recent proposals have addressed hierarchical representations (Jia et al., 2013; Lewis and Frank, 2013; Abbott et al., 2012), but they are biased to one modality or a static database.

In this chapter, we propose a deep architecture of hypernetworks for automatically constructing visual-linguistic knowledge by dynamically learning concepts represented with vision and language from videos, i.e., a deep concept hierarchy (DCH) (Ha et al., 2014a). DCH consists of two or more concept layers and one layer of multiple modalities. The concepts at the higher levels represent more abstract concepts than at the lower layers. The modality layer contains the populations of many microcodes encoding the higher-order relationships among two or more visual and textual variables (Zhang et al., 2012). Each concept layer is represented by a hypergraph. This structure coincides with the grounded theory of the human cognition system where a concept is grounded in the modality-specific regions (Kiefer and Barsalou, 2012). The structure enables the multiple levels of concepts to be represented by the probability distribution of the visual-textual variables.

The concept construction of DCH from videos involves two technical issues. One is to search a huge space of DCH represented by hypergraphs. The other is to deal with concept drift contained in the video data. For handling these two issues, DCH uses a method based on a Monte-Carlo simulation for efficiently exploring the search space, i.e., a graph Monte-Carlo (graph MC). The graph MC is a stochastic method for efficiently finding desired graph structures by the repetition of probabilistically generating connections among nodes using observed data instead of sampling. The model structure flexibly grows and shrinks by the graph MC, in contrast to other deep learning models. DCH incrementally learns the concepts by the graph MC and the weight update process while observing new videos, thus robustly tracing concept drift and continuously accumulating new conceptual knowledge. This process is formalized as a sequential Bayesian inference. The learning mechanism is inspired by the cognitive developmental process of children constructing the visually grounded concepts from multimodal stimuli (Meltzoff, 1990).

For evaluation, we used the whole collection of cartoon videos for children, entitled ”Pororo”, consisting of 183 episodes with 1,232 minutes of playing time. Experimental results show DCH faithfully captures visual-linguistic concepts at multiple abstraction levels, reflecting the concept drift in the progress of the stories. Technically, we investigate the effective combinations of hierarchy architectures and graph MC variants to construct the DCH fast, flexibly, and robustly based on sequentially observed data over an extended period of time. We also present the application of the concept hierarchies for story- and context-aware conversion between the video scenes and the text subtitles.

4.2 Visual-Linguistic Concept Representation of Catoon Videos

To be concrete, we start with the video data from which we extract the vision-language concepts. The whole data set used in this chapter consists of episodes, which are preprocessed into the sequences of sentence-image pairs by capturing a scene when a subtitle appears. The vocabulary for the visual words is defined by SIFT, RGB color and MSER features. If we represent a textual word as w_i and the visual word as r_i, the utterance-scene is represented as a vector of the form:

$$\mathbf{x}^{(t)} = (\mathbf{w}^{(t)}, \mathbf{r}^{(t)}) = (w_1, ..., w_{|\mathbf{w}^{(t)}|}, r_1, ..., r_{|\mathbf{r}^{(t)}|}) \tag{4.1}$$

$$D_N = \{(\mathbf{w}^{(t)}, \mathbf{r}^{(t)}) | t = 1, ..., T\} \tag{4.2}$$

Figure 4.1 shows four instances of scene-utterance pairs transformed from cartoon videos. The objective is to construct a knowledge representation from the data that keeps main conceptual information.

4.3 Deep Hypernetworks for Modeling Visual-Linguistic Concepts

We address an extension of multimodal hypernetworks for learning concepts represented by vision and language in this section. The proposed model uses multiple layeres of hypernetworks, which represent concept hierarchies composing video stories represented as visual and textual variables i.e., deep concept hierarchy (DCH). DCH consists of two kinds of layers; multiple concept layers and a sparse code layer. The multiple concept layers include one or more layers of concept variables. Variables in higher layers represent more abstract concepts and nodes in lower layers characterize more concrete ones. The sparse code layer involves a large

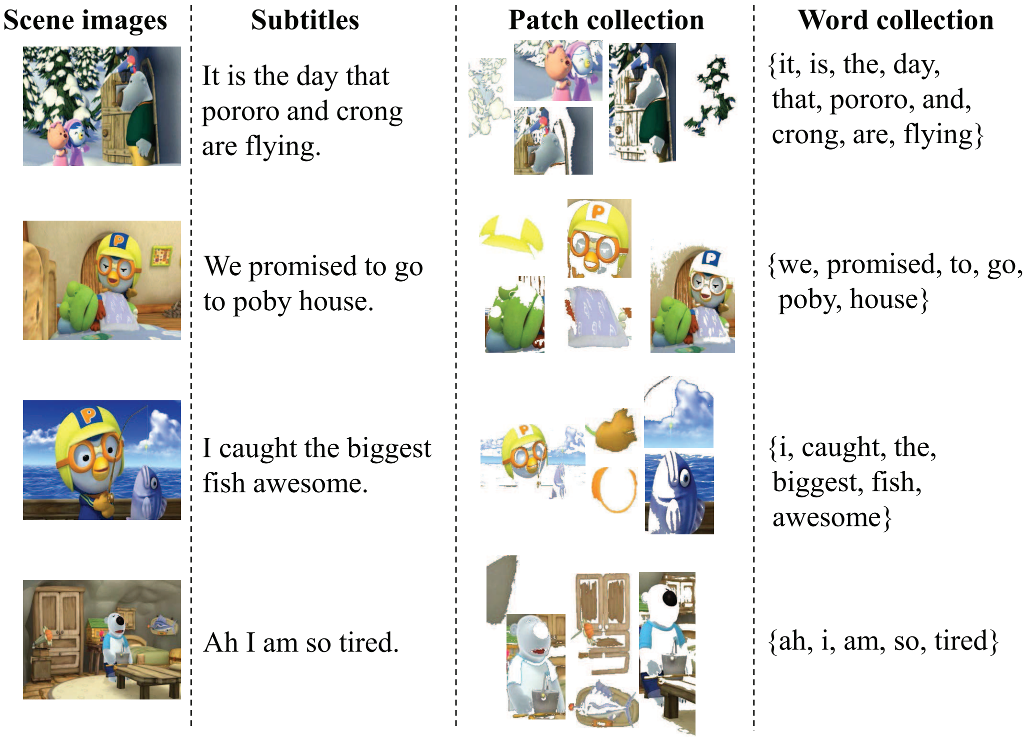

Scene images
Subtitles
Patch collection
Word collection
It is the day that pororo and crong are flying.
{it, is, the, day, that, pororo, and, crong, are, flying}
We promised to go to poby house.
{we, promised, to, go, poby, house}
I caught the biggest fish awesome.
{i, caught, the, biggest, fish, awesome}
Ah I am so tired.
{ah, i, am, so, tired}

Figure 4.1: Examples of utterance-scene pairs extracted from cartoon videos

population of many multimodal microcodes. A microcode contains two or more visual and textual variable values, thus encoding a small association rule between two modalities. Because the microcode represents a small subpattern among very large spaces of feature combination, the layer including the population is called sparse code layer. This structure not only allows the model to represent various levels of concepts involved in the video contents but also enables the concepts to be characterized with the probability distribution of visual-textual variables. Figure 4.2 illustrates an architecture of a deep hypernetwork for characterizing the hierarchy of visual-lingustic concepts.

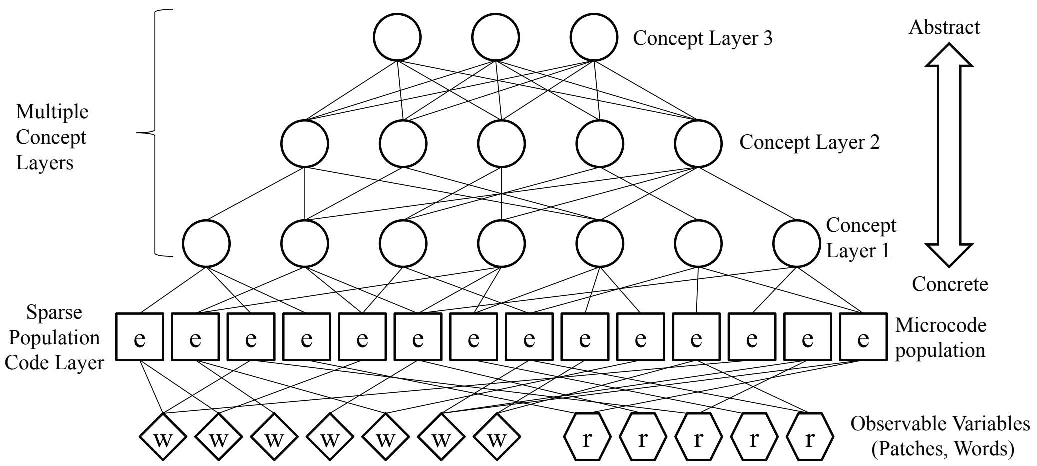


Figure 4.2: Architecture of deep hypernetwork for representing concept hierarchies

4.3.1 Sparse Population Coding

Sparse population coding (SPC) proposed by Zhang et al. is a method for representing and learning concepts from dynamic data, using a large population of multimodal information chunks (Zhang et al., 2012). The SPC simulates a situated word learning from cartoon videos in childhood and this process is addressed in terms of Bayesian inference. In specific, SPC is a principle to encode data of n variables compactly using multiple subsets of size k. The subset is called a microcode and, typically, its size is small, i.e. $k << n$, and thus sparse. To deal with concept drift contained in the video stories, SPC is based on a population coding scheme suitable for incremental learning. The population consists of large number of microcodes encoding the relationships between visual and textual variables. In a SPC model, the concepts are implicitly represented with subsets of microcode population and this means that the empirical distribution of the concepts is characterized in the form of a finite mixture with the population.

Figure 4.3 shows an example of the concept representation in SPC. In Figure 4.3, each h denotes a microcode, which is represented with a set of two or more words and image patches. Formally, a microcode population characterizes the empirical distribution of the concepts in the form of a finite mixture. When data $x = (w, r)$ are given, the distribution can be defined:

$$P(\mathbf{x}|\theta) = \sum_{i=1}^{M} \alpha_i f_i(\mathbf{x}|e_i) \tag{4.3}$$

where e_i and α_i denote a microcode and its weight, and $f_i(\mathbf{x}|e_i)$ is a density function. Also, α_i satisfies the following conditions:

$$0 \leq \alpha_i \leq 1 \;\; and \;\; \sum_{i=1}^{M} \alpha_i = 1 \tag{4.4}$$

In above equation, a model parameter θ is defined as $\theta = (\alpha, e)$, where e and α are the sets of M microcodes and the weights associated with the microcodes. Then, the empirical distribution of the observed video data consisting of continuous T scenes $D = \{\mathbf{x}^{(t)}\}_{t=1}^{T} = \{(\mathbf{w}, \mathbf{r})^{(t)}\}_{t=1}^{T}$ can be represented by the population code:

$$P(D|\theta) = \prod_{t=1}^{T} P(\mathbf{x}^{(t)}|\theta) = \prod_{t=1}^{T} \sum_{i=1}^{M} \alpha_i f_i(\mathbf{w}^{(t)}, \mathbf{r}^{(t)}|e_i) \tag{4.5}$$

The representation of SPC model enables the concepts to be implicitly characterized by the sparse subpatterns between visual and textual variables.

SPC uses a flexible hypergraph structure as the model representation for handling concept drift in the increasing multimodal data. A SPC model can be equivalently transformed into a hypernetwork when we consider the sets of image patches and textual words as a vertex set. Also, each microcode is converted into a clique connecting the vertices encoding the patches and the words of the microcode. Figure

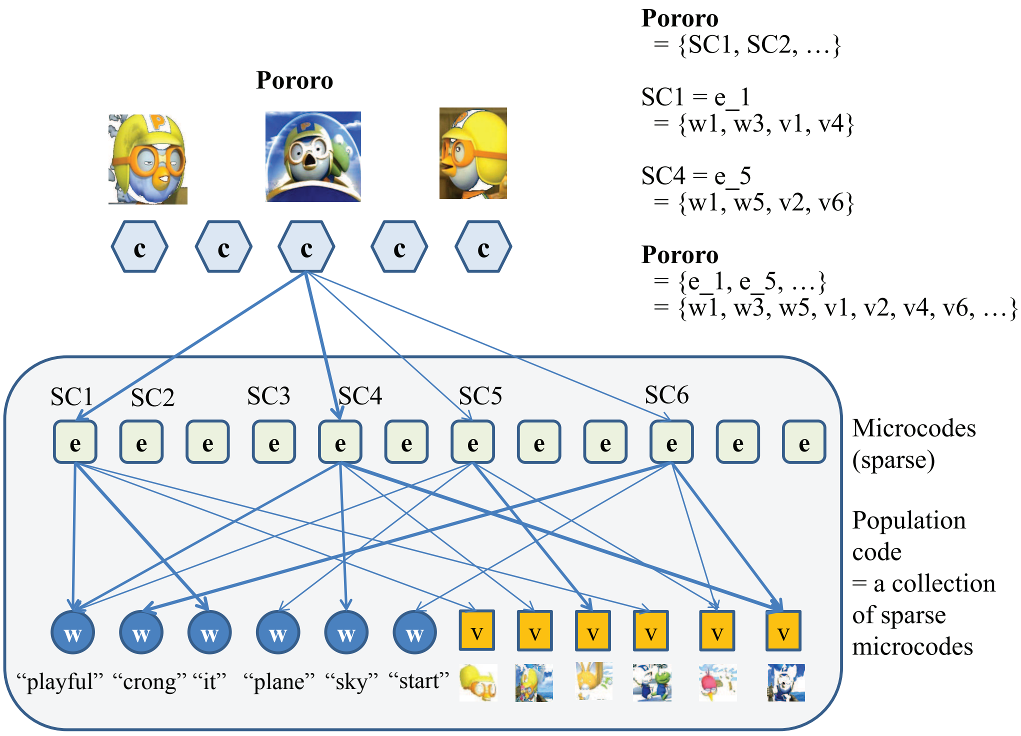
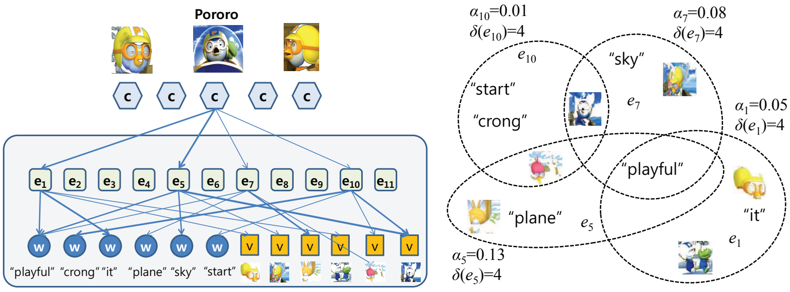


Figure 4.3: An example of the concept representation of Pororo via sparse population coding. SC denotes a sub-concept corresponding to a microcode

4.4 shows an example of the hypernetwork representation of a SPC model. Therefore, a SPC model implicitly characterizes the concepts in video with higher-order associations between image and text variables.

4.3.2 Deep Hypernetworks for Concept Hierarchies

SPC can be considered as a hypergraph, where the hyperedges represent the microcodes. An equivalent representation is a two-layer network where the upper-layer nodes indicate microcodes (hyperedges) and the lower-layer nodes indicate



(a) SPC model

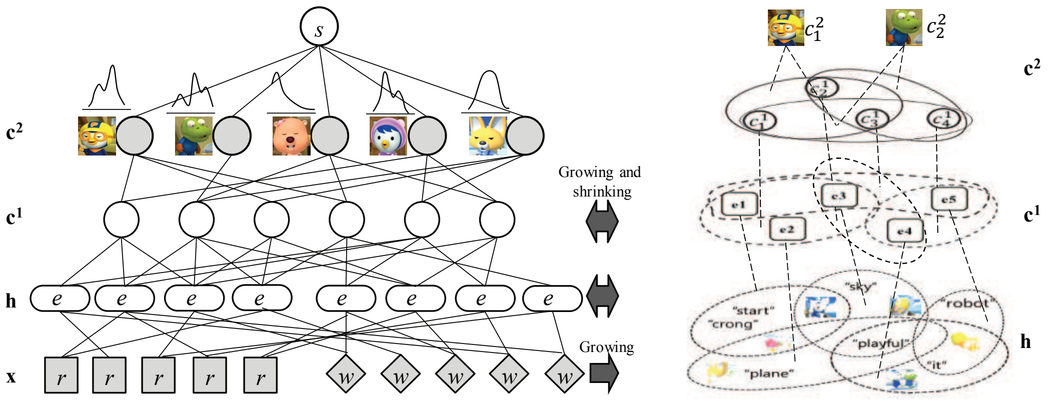
(b) Hypergraph representation

v(*w*)	*d*(*v*)	*v*(*r*)	*d*(*v*)
playful	0.26		0.05
crong	0.01		0.08
it	0.05		0.13
plane	0.13		0.05
sky	0.08		0.14
start	0.01		0.09

(c) Degree of vertices

Figure 4.4: An example of sparse population coding model and its hypernetwork representation

the data variables. Though the representation power is large, the number of upper-layer nodes may grow fast with the growing number of input units, i.e. the visual and textual vocabulary sizes in our video data. To resolve this problem, we introduce additional layers, resulting in a deep concept hierarchy (DCH). DCH is a model representing the hierarchy of visually grounded concepts, using multimodal sparse population coding. Dissimilar to SPC models where concepts are implicitly represented and learned, concept variables are introduced into the model in DCH. To representing diverse levels of concepts, the model contains one or more concept layers. Nodes in a higher layer encode more abstract concepts and those of a lower layer characterize more concrete concepts. A node in a layer can connect to nodes in adjacent layers of the layer, and the nodes of the most specific conceptual layer can connect to a subset of the population of sparse codes which encode the association between observable multimodal variables. The connections between layers of a DCH are sparse, which is contrasted to the deep neural networks that have full connectivity between layers. This sparse and hierarchical structure reduces



(a) Example of a deep concept hierarchy

(b) Hypergraph representation of (a)

Figure 4.5: An Example of deep concept hierarchy with two conceptual layers learning from cartoon videos. Abstract concept variables denote the characters appearing in the videos. Gray boxes are observable variables.

the model complexity and DCH pursues a parse modular hierarchical structure, as found in human brains.

Mathematically, DCH represents the empirical distribution of data using a multiple layers of microcodes or concepts. Consider a DCH model with two concept layers. Assume that a node of the top concept layer denotes a character appearing in the video. Let c^1=$(c^1_1, ..., c^1_{K_1})$, c^2=$(c^2_1, ..., c^2_{K_2})$, $\mathbf{w} = (w_1, ..., w_M)$, and $\mathbf{r} = (r_1, ..., r_N)$ denote the binary vectors representing the presence of concrete and abstract concepts, textual words, and image patches in the scene-text pair, where K_1, K_2, M, and N is the number of two concepts, the size of word vocabulary, and the size of the patch dictionary. M and N increase whenever observing new words and patches. The probability density of a scene-text pair for a given $\mathbf{h} = (\mathbf{e}, \alpha)$, $\mathbf{c^1}$, and $\mathbf{c^2}$ can be formulated as follows:

$$P(\mathbf{r}, \mathbf{w}|\mathbf{c}^1, \mathbf{c}^2) \ = \sum\nolimits_h P(\mathbf{r}, \mathbf{w}|\mathbf{h}, \mathbf{c}^1, \mathbf{c}^2)P(h|\mathbf{c}^1, \mathbf{c}^2) \tag{4.6}$$

where e and α denote the population of microcodes and their weights. Each microcode e is defined as two sparse binary vectors whose size is M and N at the time when the scene is observed, respectively. Therefore, DCH can model the concepts as probabilistic associations among words and images. Figure 2 (b) shows an instance of DCH with two concept layers learning concepts from videos. In addition, similar to deep networks, the joint probability of all the layer nodes is computed by the production of the conditional probabilities:

$$P(\mathbf{x}, \mathbf{h}, \mathbf{c}^1, \mathbf{c}^2) = P(\mathbf{x}|\mathbf{h})P(\mathbf{h}|\mathbf{c}^1)P(\mathbf{c}^1|\mathbf{c}^2) \tag{4.7}$$

Deep hypernetworks have some advantages in model structure for learning from continuously increasing and nonstationary data, compared to conventional multimodal learning models including topic models and deep learning models. Table 4.1 describes pros and cons of deep hypernetworks over topic models and deep learning models. As shown in Table 4.1, deep hypernetworks have flexible model structures suitable for incrementally learning from data. In addition, by using hypergraphs as a model representation and denoting semantic variables such as image patch and textual words as nodes, they can be used a efficient method for representing knowledge of the contents characterized with the data. Furthermore, the sparse structure of the models reduces the model complexity. However, dissimilar to models using a fixed structure, deep hypernetworks require a structure learning method which has large influence on the model performance.

Table 4.1: Comparison of deep hypernetworks to other multimodal leaning models

Criteria	Topic models	Deep networks	Deep HNs
Model structure	Fixed	Fixed	Flexible
Observable node	Fixed	Fixed	Flexible
Semantic representation	Middle	Low	High
Readability of hidden nodes	Low	Middle	High
Incremental learning	Difficult	Middle	Suitable
Model complexity	Middle	High	Low
Knowledge representation	Low	Low	High
Structure learning	No	No	Yes
Learning strategy	Stochastic	Gradient descent	Stochastic
Node value	Cont./Disc.	Cont./Disc.	Discrete
Dynamic learning	High	Low	Low

4.3.3 Implication of Deep Hypernetworks on Cognitive Modeling

A hypernetwork was proposed as a simulation model for DNA molecular computing (Zhang and Jang, 2005b), using the population of many small memory chunk. The representation of hypernetworks is similar to the sparse and population representation of human brains where information is stored and processed to be distributed (Quiroga, 2012), thus being suitable for characterizing human cognitive process. In particular, the introduction of a hierarchy of multiple hypernetwork layers and multimodal encoding enables the model to be used as a cognitive model based on the grounded cognition theory of the human conceptual systems (Kiefer and Barsalou, 2012). In the grounded cognition theory, cognition is grounded in the modality-specific systems of brains for the perception and the body for action. In

DCH that is an instance of deep hypernetworks for concept modeling, learned high-level abstract concepts are connected to visual and lingustic variables and this can be considered as a architecture grounded in the modality-specific systems despite the absence of the explicit body-action systems in the model. However, the model can extended into a action-grounded system by implementing the model as the intelligence systems of robots. In addition, main themes associated with the grounded conceptual systems including simulation, emergence, dynamic representation, and situatedness can be handled by deep hypernetworks. By inference of the learned model, many abstract and concrete concepts are ignited for the given external stimuli and the emerged concepts are represented as vision and language through the probabilistic simulations. In addition, whlie observing new videos, concepts are incrementally learned and the concept representations dynamically changes considering the constructed concept hierarchies. Therefore, deep hypernewtorks are a computational model suitable for cognitive modeling in the viewpoint of the grounded cognition theory of the conceptual systems.

4.4 Learning of Deep Hypernetworks

4.4.1 Problem Space of Deep Hypernetworks

DCH uses a hierarchy of the hypergraph structure as the model representation. Each layer of a DCH model can be equivalently transformed into a hypergraph as shown in Figure 4.5. However, when a hypergraph and a conventional graph have the same vertex set, the number of possible edges of a hypergraph is much more than one of a graph due to the definition. For a *k*-hypergraph, the number of possible hyperedges are

$$|\mathbf{e}| = C(n,k) = \frac{n!}{k!(n-k)!} \tag{4.8}$$

where $n = |V|$ and $C(n,k)$ denote the number of cases to choose k items from a set with n elements. When Ω is denoted as the set of all the hypergraphs, the size is equal to

$$|\Omega| = 2^{C(n,k)} \tag{4.9}$$

Therefore, the number of possible hyperedges with degree from zero to n and the size of the space of $(0,n)$-hypergraphs are

$$|\mathbf{e}| = \sum_{k=0}^{n} C(n,k) = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} = 2^n \;\; and \;\; |\Omega| = 2^{\kappa \cdot 2^n} \tag{4.10}$$

where κ is the maximum number of duplicates of a hyperedge in a hypergraph. Then the problem space of a DCH model is equal to $2^{\kappa \cdot 2^n}$, where $|\mathbf{x}|$ denotes the size of the observable variable set, since each node connection of the conceptual layers can be transformed into a corresponding hyperedge containing a subset of x. It is infeasible to explore this huge search space with an exhaustive or a gradient strategy. For efficiently searching the huge space, we propose a stochastic method, i.e., graph Monte-Carlo (graph MC).

4.4.2 Graph Monte-Carlo Simulation

We propose a method for efficiently constructing hypergraphs incrementally from incoming data. The idea is to use Monte Carlo search on the hypergraph space. The resulting graph Monte-Carlo method (Graph MC) assumes two conditions:

i) The graph structure in the t-th iteration is determined by that of the t-1 the iteration.

ii) Estimating the empirical distribution asymptotically converges to exploring all theoretical spaces when data are large enough.

Formally, for a given dataset D, an optimal hypergraph G^* corresponding to a model is formulated with Bayes rule:

$$G* = \arg\max_{G_t} P(G_t|D) = \arg\max_{G_t} P(D|G_t)P(G_{t-1}) \tag{4.11}$$

where G_t is a k-hypergraph in the t-th time step. G is constructed to maximize $P(G_t|D)$ by the repetition of replacing hyperedges whenever observing the data:

$$\Delta G = \Delta G \cup \{e\} \quad and \quad e = \bigcup_{m=1}^{k} v(x) \tag{4.12}$$

where ΔG is a new hyperedge set, and e and $v(x)$ denote a generated hyperedge and the vertex corresponding to a variable x. Both the initial values of ΔG and e are empty. $P(e)$ denotes the probability with which e is generated. The graph MC is addressed in terms of the Metropolis-Hastings algorithm (Newman and Barkema, 1999) under two conditions:

i) A hypergraph G is factorized by its hyperedges to represent a probability distribution (Besag 1974).

ii) ΔG is generated to equivalently represent a sampling instance x.

Then, G^* representing the empirical distribution of the observed data can be constructed by the graph MC. The learning strategy of the graph MC is determining the probability of which vertices are connected as hyperedges, $P(v(x))$. Note that $P(v(x))$ is computed from the currently observed data instance according to assumption ii). We define $P(v(x))$ based on three different approaches.

Uniform graph Monte-Carlo

Uniform graph Monte-Carlo (UGMC) uses the same probability as $P(v(x))$ for all the variables with the positive value of the data. Then, the probability is defined as follows:

$$P(v(x)) = \frac{1}{\left|\{x|x \in x_+^{(n)}\}\right|} \tag{4.13}$$

$$P(e) = \frac{1}{C(k, |x_+^{(n)}|)} \tag{4.14}$$

where $x_+^{(n)}$ denotes the set of variables with the positive value of the n-th data instance. Then, all the possible hyperedges for a given instance are generated with the same probability.

Poorer-richer graph Monte Carlo

The $P(e)$ of each possible hyperedge for a given instance is different from each other in poorer-richer graph Monte-Carlo (PRGMC). In PRGMC, a vertex more included in a hypergraph has higher probability. The $P(v(x))$ of PRGMC is defined as follows:

$$P(v(x)) = \frac{R^+ \{d(v(x))\}}{|x|} \tag{4.15}$$

where $R^+(.)$ is a rank function in ascending order, $d(v)$ is the degree of vertex of v. For enabling new variables not existing in G_{t-1} to be selected, their $d(v)$ is set to a small value. This approach makes a hypergraph contain the patterns which frequently appear in the training data. Therefore, PRGMC constructs a smaller and denser hypergraph, compared to that built by UGMC.

Poorer-richer graph Monte Carlo

Fair graph Monte Carlo Fair graph Monte Carlo (FGMC) prefers the subpatterns less frequently appearing in the training data, contrary to PRGMC. The P(v(x)) is defined as:

$$P(v(x)) = \frac{R^{-}\{d(v(x))\}}{|x|} \tag{4.16}$$

where $R^{-}(.)$ is a rank function in descending order. Therefore, a larger and sparser graph is constructed by FGMC and the concepts are represented with much more diverse words and patches.

4.4.3 Learning of Concept Layers

To learning the concept layers we should address three issues: i) determining the number of the nodes of the concrete concept layer $\mathbf{c^1}$ ($\mathbf{c^1}$-nodes), ii) associating between $\mathbf{c^1}$-nodes and modality layer $\mathbf{h}$, and iii) associating between $\mathbf{c^1}$-nodes and the abstract concept nodes ($\mathbf{c^2}$-nodes). The idea is to split the hyperedge set in h into multiple subgraph clusters, which correspond to the nodes of the c^1 layer. The number of the $\mathbf{c^1}$-nodes are determined based on the distribution of the mean similarities among the hyperedges of a subgraph on all the clusters:

$$Sim(\mathbf{h}^m) = \frac{Dist(\mathbf{h}^m)}{|\mathbf{h}^m|} \tag{4.17}$$

where $\mathbf{h}^m$ denotes the subgraph associated with the m-th $\mathbf{c^1}$-node and $Dist(\mathbf{h}^m)$ is the sum of the distance between all the hyperedges of $\mathbf{h}^m$. Then, the distance is estimated by converting the words into the real-value vectors by word2vec (Mikolov et al., 2013b). If $Sim(\mathbf{h}^m) > \theta_{max}$, $\mathbf{h}^m$ is split into two subgraphs and a new $\mathbf{c^1}$-node is added into the c^1 layer and associated with one of the split subgraph. On the other hand, if all the mean similarities are smaller than θ_{min}, the number is reduced and the associations are conducted again. θ_{max} and θ_{min} are adaptively determined from the mean and the variance of the similarity.

The connectivity between two concept layers are determined by the constitution of hyperedges of the subgraph associated with c1-nodes. Each hyperedge includes the information on the character appearance of the scene from which it is generated. Then, a $\mathbf{c}^1$-node connects $\mathbf{c}^2$-nodes corresponding to characters which is included in hyperedges that the subgraph of the $\mathbf{c}^1$-node contains. That is, $\mathbf{c}^2$-nodes are associated with a $\mathbf{c}^1$-node when the characters corresponding to the $\mathbf{c}^2$-nodes appear in the hyperedges of the subgraph associated with the c1-node. The weight of the connection is defined by the weighted ratio of each character appearance in the microcode cluster:

$$\omega(c_i^1, c_j^2) = \frac{\sum_{h_m \in \mathbf{h}^i} \alpha_m C(c_j^2, h_m)}{\sum_{h_m \in \mathbf{h}^i} \alpha_m} \tag{4.18}$$

where $C(c_j^2, h_m)$ is the indicator function that yields 1 when the character corresponding to the j-th node of $\mathbf{c}^2$-layer, c_j^2, appears in the scene from which the m-th microcode is generated.

4.4.4 Incremental Concept Construction

DCH learns incrementally, i.e. builds the visual-linguistic concepts dynamically while sequentially observing scene-text pairs. We use all the scene-text pairs of one episode as a mini corpus. On sequential observation of the episodes, DCH predicts the concepts from the population and updates the population from the observed data and characters. Formally, this implements a sequential Bayesian estimation:

$$P_t(\mathbf{h}, \mathbf{c}^1 | \mathbf{r}, \mathbf{w}, \mathbf{c}^2) = \frac{P(\mathbf{r}, \mathbf{w} | \mathbf{h}, \mathbf{c}^1, \mathbf{c}^2) P(\mathbf{c}^2 | \mathbf{c}^1, \mathbf{h}) P_{t-1}(\mathbf{h}, \mathbf{c}^1)}{P(\mathbf{r}, \mathbf{w}, \mathbf{c}^2)} \tag{4.19}$$

where P_t is a probability distribution at the t-th episode. When observing the t-th episode, the prior distribution is updated to the posterior distribution by calculating the likelihood and normalizing. Then, the posterior is used as the prior for learning from the next episode. Note that the $P(\mathbf{r}, \mathbf{w}, \mathbf{c}^2)$ is independent on the model because $(\mathbf{r}, \mathbf{w})$ and $\mathbf{c}^2$ are given from the observed data. Therefore, 4.19 is reformulated when the empirical distributions are used:

$$P_t(\mathbf{h}, \mathbf{c}^1|\mathbf{r}, \mathbf{w}, \mathbf{c}^2) \propto \prod_{d=1}^{D_t} \left\{P(r^{(d)}, w^{(d)}|\mathbf{h}, \mathbf{c}^1, \mathbf{c}^2)P(\mathbf{c}^2|\mathbf{c}^1)P(\mathbf{c}^1|\mathbf{h})P_{t-1}(\mathbf{h})\right\}. \tag{4.20}$$

The data generation term is divided into textual and visual features:

$$\log P(r^{(d)}, w^{(d)}|\mathbf{c}^2, \mathbf{c}^1, \mathbf{h}) = \sum_{n=1}^{N} \log P(r_n^{(d)}|\mathbf{c}^2, \mathbf{c}^1, \mathbf{h}) + \sum_{m=1}^{M} \log P(w_m^{(d)}|\mathbf{c}^2, \mathbf{c}^1, \mathbf{h}). \tag{4.21}$$

Then the probability that the m-th element of the word vector is 1 is defined as follows:

$$P(w_m^{(d)} = 1|\mathbf{c}^2, \mathbf{c}^1, \mathbf{h}) = \exp\left(s_m^w - \sum_{i=1}^{|\mathbf{e}^c|} \alpha_i\right), \tag{4.22}$$

$$P(r_n^{(d)} = 1|\mathbf{c}^2, \mathbf{c}^1, \mathbf{h}) = \exp\left(s_n^r - \sum_{i=1}^{|\mathbf{e}^c|} \alpha_i\right), \tag{4.23}$$

s.t. $s^w = \sum_{i=1}^{|\mathbf{e}^c|} \alpha_i e_i^w$ and $s^r = \sum_{i=1}^{|\mathbf{e}^c|} \alpha_i e_i^r$
where s_m is the m-th value of s and $\mathbf{e}^c$ denotes the subpopulation of microcodes associated with c^1. $e_i{}^w$ and $e_i{}^r$ denotes the textual and visual vectors of the i-th microcode. The second term of 4.21 is related to predicting the characters from the mixtures of concrete concepts. It is defined to prefer more distinct concrete concepts for each character variable. The third term reflects the similarities of the subpopulation for each concrete concept node. The last term is determined from the used strategy of the graph MC.

The initial weight of the microcodes is defined as a function of how frequently the words and patches of the microcode occur in the observed data:

$$\alpha_i = \sum_{d=1}^{D} \Big\{ g(e_i) f(r^{(d)}, w^{(d)}; e_i) \Big\}, \tag{4.24}$$

s.t.

$$f(r^{(d)}, w^{(d)}; e_i) = \begin{cases} 1, \text{ if}\Big(r^{(d)} \cdot e_i^r + w^{(d)} \cdot e_i^w \Big) \Big/ e_i e_i^{\mathrm{T}} > \kappa \\ \quad\quad 0, \text{otherwise} \end{cases}$$

where $r^{(d)} \cdot e_i^r$ and $w^{(d)} \cdot e_i^w$ denote the inner product of the textual and the visual vectors of the i-th microcode between and , respectively. $g(e_i)$ is the geometric mean of the *TF-IDF* values of the words with 1 in e_i, and this term prevents the abnormal large weight of a microcode containing functional words only. κ is a nonnegative constant less than 1. Whenever observing a new episode, it is updated:

$$\alpha_i^t = \lambda \alpha_i + (1 - \lambda) \alpha_i^{t-1} \tag{4.25}$$

λ in 4.25 is a constant for moderating the ratio of the new observed episode and the previous episodes.

4.5 Incremental Concept Construction from Catoon Videos

4.5.1 Data Description and Parameter Setup

We use cartoon videos, called "Pororo", of 14 DVD titles with 183 episodes and 1,232 minutes of playing time. By preprocessing, each scene is captured whenever a subtitle appears, transforming all the videos into the set of 16,000 utterance-scene pairs. A scene image is represented by a bag of image patches extracted by maximally stable external regions (MSER) (Matas et al., 2004) and each patch is defined as a feature vector using SIFT (Lowe, 2004) and RGB color. We used a DCH model with two concept layers. A microcode consists of two image patches and a

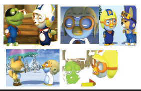
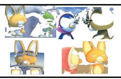
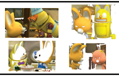
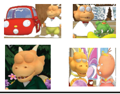
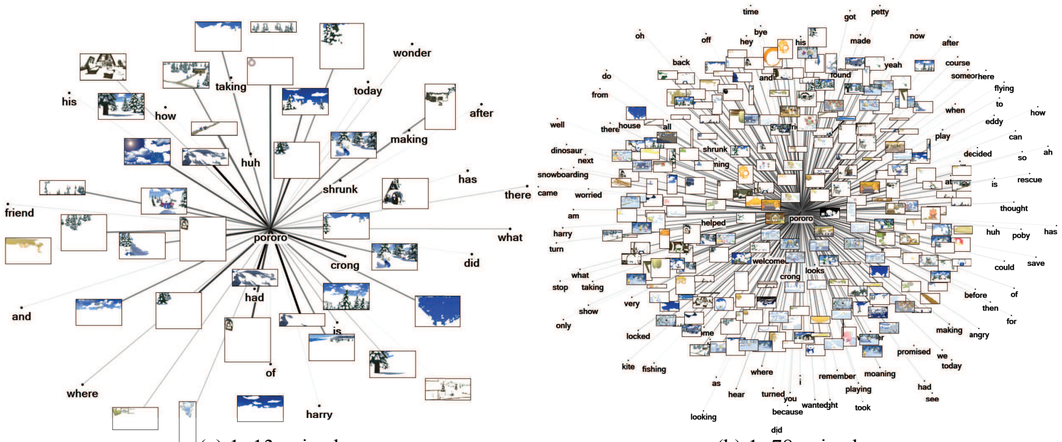
	1~13 episodes (1 DVD)			1~183 episodes (14 DVDs)		
Concepts	Visual nodes	# of nodes (V/L)	Top 15 linguistic nodes	Visual nodes	# of nodes (V/L)	Top 15 linguistic nodes
Pororo		986/230	crong, you, clean, over, draw, huh, to, it, I, up, said, the, moving, is, pororo		12870/1031	crong, you, snowboarding, transforming, rescuing, pororo, the, lamp, seven, are, quack, yellow, not, lollipop, cake,
Eddy		644/198	I, ear, art, midget, game, nothing, say, early, diving, lost, middle, lesson, case, because, snowballs		9008/860	transforming, I, hand, careful, throw, art, suit, midget, farted, reverse, stage, luggage, gorilla, pole, cannon
Tongtong	-	0/0	-		1812/429	kurikuri, doodle, doo, avoid, airplane, crystal, puts, branch, bland, finding, pine, circle, kurikuritongtong, bees, talent

Figure 4.6: Visual-linguistic representation and development of three character concepts of video contents. A scene-utterance pair is represented by the sets of image patches and words and the concepts of the video stories are represented by these patches and words. *Tongtong* is not seen in episodes 1-13 and appears in episode 56 for the first time.

phrase with three consecutive words. The image patches are selected by UGMC and a phrase is selected with the maximum value of $P(v(x))$ of the words in the phrase. The initial number of c^1-nodes starts at 10 and θ_{max} and θ_{min} are defined as follows:

$$\theta^t_{\mathrm{max}} = \begin{cases} \mu^t + \eta \cdot (\mu^t - \mu^{10}) \cdot \sigma^t, \ t > 10 \\ \mu^t + \eta \cdot \sigma^t, \ \ t \leq 10 \end{cases} \quad , \ \theta^t_{\mathrm{min}} = 0 \tag{4.26}$$

where μ^t and σ^t denote the mean and the standard deviation of the subgraph similarities after observing the t-th episode, and is a constant for moderating the increasing speed of the c^1 layer size. Lager η reduces the speed and we set it to 0.75.

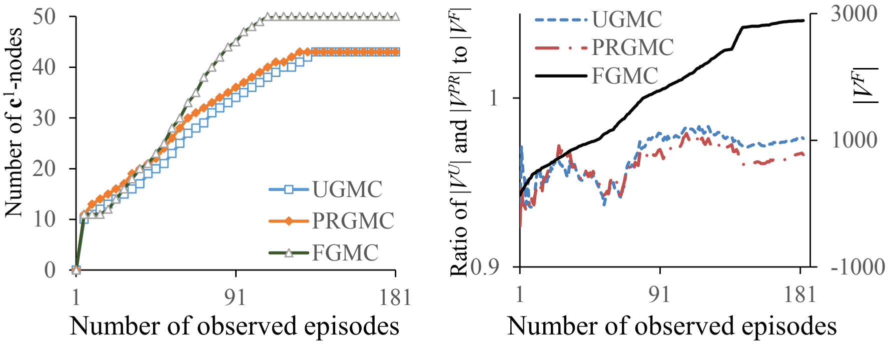


(a) 1~13 episodes (b) 1~78 episodes

Figure 4.7: Evolution of *Pororo*'s visual-lingustic concept maps

4.5.2 Concept Representation and Development

To demonstrate the evolution of concepts in DCH, we have examined how the characters, such as *"Pororo"*, *"Eddy"*, and *"Tongtong"*, are differently described as the story unfolds. Figure 1 compares the descriptions after learning up to episode 13 (DVD 1) and 183 (DVD 14). Considering the fact that *Pororo* is a brother of *Crong*, *Tongtong* casts *"Kurikuri"* for magic, and *Eddy* is an engineer, the descriptive words for each character are suitable. We observe that the number of visual and linguistic nodes tends to increase. This is because the concepts continuously develop while observing the videos. However, we indicate that the textual nodes representing each character concept does not linearly increase but saturates as the increment of the amounts of observed videos, comparing the number of nodes of 1 DVD to that of 14 DVDs. This is caused by that new words seldom appear in later episodes. The saturation of the text nodes enables our method to scale up to much more video



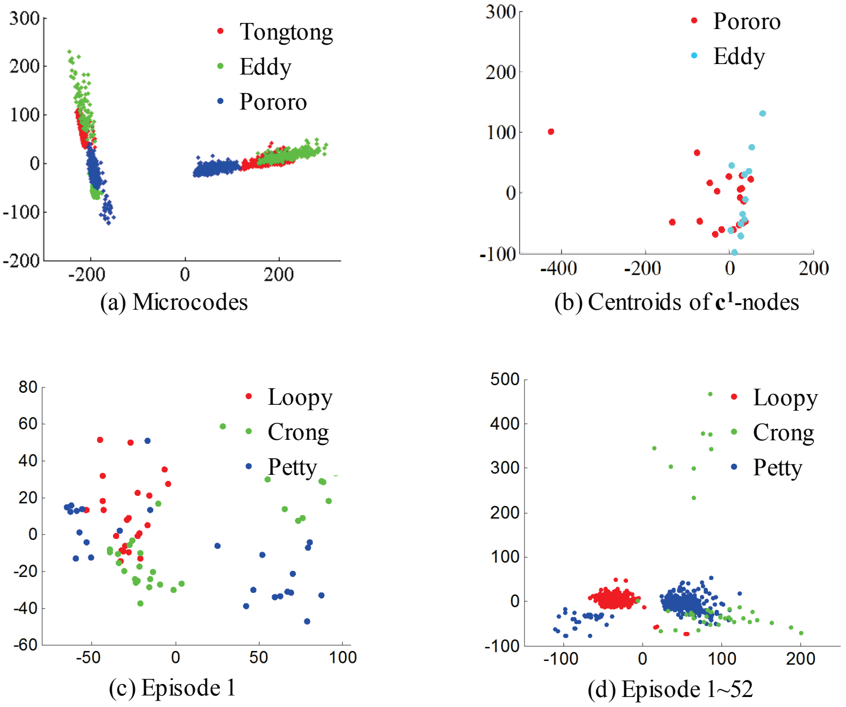
(a) Nodes of the $\mathbf{c}^1$-layer (b) Nodes of textual word variables

20

Figure 4.8: Changes of model complexity according to the learning strategies of the graph MC. In (b), VU, VPR, and VF denote the vertex sets of the model constructed by UGMC, PRGMC, and FGMC

data. Figure 4.7 illustrates the concept evolution of Pororo as the observed videos increase. As shown in Figure 4.7, *Pororo* concept becomes much more complex by being associated with more words and patches.

Specifically, we observed that the number of $\mathbf{c}^1$-nodes increases in early stages and then saturates, in addition to textual nodes. Figure 4.8(a) shows the change of the number of $\mathbf{c}^1$-nodes as the observed episodes increase. Regardless of the learning strategy of the graph MC, the number of $\mathbf{c}^1$-nodes fast increases in early episodes and then saturates after that. This indicates that new concrete concepts are learned rather earlier and, as time goes on, familiar concepts reappear. In addition, Figure 4.8(a) compares the complexity growth curves of DCH by three learning methods. FGMC is the most fast-growing strategy employing more c1-nodes because it tends to select diverse words and patches, as compared to UGMC and PRGMC. This is verified by Figure 4.8(b) which shows more vertices are included in the models



(a) Microcodes

(b) Centroids of $\mathbf{c}^1$-nodes

(c) Episode 1

(d) Episode 1~52

Figure 4.9: PCA plot of microcodes associated with the concrete concept nodes (c1-nodes) and their centroids of the models learned from 183 episodes by UGMC.

constructed by FGMC.

To see if DCH correctly learned the distinguishable concepts, we have analyzed the $\mathbf{c}^1$-nodes by PCA. Figure 4.9 shows that different characters are well discriminated by the learned microcodes (the first component ratio = 0.70 in (a)). This indicates that a $\mathbf{c}^2$-node corresponding to a character is associated with $\mathbf{c}^1$-nodes distinguishably representing the property of the character.

4.5.3 Character Classification via Concept Learning

The learned concepts via deep concept hierarchies can be used for classifying the characters appearing in scenes. We present the character classification as a quantitative performance of the proposed concept model. The character classification is defined to classify which characters appear from the given scene images and subtitles. Because the class label of a scene is represented by not a single value but a binary vector of whether each character appears in the scene, this problem belongs to a multidimensional classification task (Zhang and Zhou, 2006). Multidimensional classification is formalized as follows:

$$h : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \rightarrow \Omega_{C_1} \times \cdots \times \Omega_{C_d}, \tag{4.27}$$

$$(x_1, ..., x_m) \mapsto (c_1, ..., c_m), \tag{4.28}$$

where C_i and X_j for all i=1, ..., d and j=1, ..., m are discrete, and Ω_{X_i} and Ω_{C_j} are sample spaces.

The used mDBMs have two hidden layers for each modality and one joint hidden layer, and the classification is carried out in the joint hidden layer using regression of each node values. Also, we used the Bayesian chain classfiers (BCCs) (Zaragoza et al., 2011), a conventional multidimensional classification method for comparison, implemented in a multilabel extension of Weka (MEKA). We used UGMC and FGMC for learning DCHs. In addition, to investigate the effect of hierarchies, we compared the results with those of SPCs. For evaluation, First 78 episodes are used as the training dataset and epsiodes from 79 to 104 are used as the test dataset. Also, we use mean average precision (MAP), a measure widely used for multiple label

Table 4.2: Character classification results

Models	BCCs	MDBMs	SPCs	UGMC	FGMC
MAP	0.314±0.012	0.630±0.007	0.553±0.021	0.612±0.035	0.643±0.032

classification:

$$MAP = \frac{1}{N}\sum_{n=1}^{N} AveP(n) \text{ and } AveP = \frac{\sum_{k=1}^{n} Precesion(k) \times rel(k)}{\#of\ relevant\ characters}, \tag{4.29}$$

where $Precision(k)$ denotes the precision at the cut-off k and $rel(k)$ is a indicator function equaling 1 when the character at rank k is a relavent character, zero otherwise. Table 4.2 shows the result of classifying the characters in the given scenes, compared to multimodal deep Boltzmann machines (mDBMs) (Srivastava and Salakhutdinov, 2012). As shown in Table 4.2, we indicate that DCHs outperform BCCs in addition to SPCs and provide competitive performance compared to MDBMs. Also, we found that FGMC showed better classification performances than UGMC and this indicate concepts learned by FGMC are represetned by more descriptive and diverse visual and textual features.

4.5.4 Vision-Language Conversion via Concept Learning

We show the result of converting between vision and language based on visual-lingustic concepts from the learned DCHs. Figure 4.10(a) shows the sentences generated from the images. The test data set consists of 183 images by randomly selecting one scene image per episode, and the results are averaged on 10 experiments. We examined how the different graph MC algorithms effect on the results.

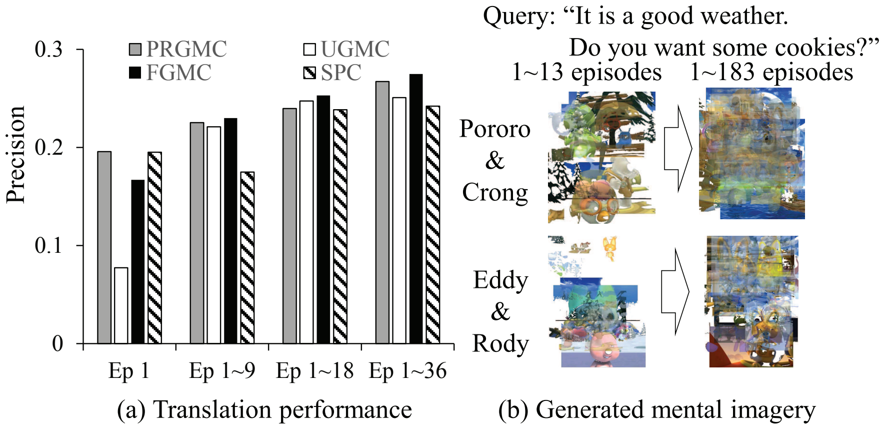


Figure 4.10: Results of vision-language conversion. precision of vision-to-language (a) and intermediate images generated from given sentences (b)

The precision of PRGMC increases faster in early videos but slower in late ones than that of FGMC. PRGMC is good at fast memorizing of main information but loses details. On the contrary, FGMC requires a more complex structure to memorize more information but shows higher accuracy. This is consistent with the results in Figure 4.8. In addition, the result shows that the introduction of concept layers improves the accuracy of the constructed knowledge. Figure 4.10(b) shows the recall of images given the text sentences. It is interesting to note that the recall images are like mental imagery as demonstrated in movie recall in humans (Nishimoto et al., 2011). Overall, the results in Figure 5 demonstrate that the more episodes the DCH learned, the more diversity are generated in sentences and images. It should be noted that this is not for free; Observing more episodes requires heavier computational costs. The tradeoff should be made by the controlling the greediness of the graph MC algorithms as examined above.

4.6 Summary

We have presented a deep concept hierarchy (DCH) for automated knowledge construction by learning visual-linguistic concepts from cartoon videos. DCH represents mutually-grounded vision-language concepts by building multiple layers of hypergraph structures. Technically, the main difficulty is how to efficiently learn the complex hierarchical structures of DCH in online situations like videos. Our main idea was to use a Monte-Carlo method. We have developed a graph MC method that essentially searches ”stochastically” and ”constructively” for a hierarchical hypergraph that best matches the empirical distribution of the observed data. Unlike other deep learning models, the DCH structure can be incrementally reorganized. This flexibility enables the model to handle concept drifts in stream data, as we have demonstrated in the experiments on a series of cartoon videos of 183 episodes.

We have analyzed and compared three strategies for the graph MC: uniform graph Monte-Carlo (UGMC), poorer-richer graph Monte-Carlo (PRGMC), and fair graph Monte-Carlo (FGMC) depending on the probability of selecting vertices. The use of hierarchy improved the generalization performance while paying slight prices in computational cost. Among the variants of the Monte Carlo algorithms, we found that the PRGMC and the FGMC work better in earlier and later stages of video observation in the visual-language translation task. Overall, our experimental results demonstrate that DCH combined with the graph MC algorithms captures the mixed visual-linguistic concepts at multiple abstraction levels by sequentially estimating the probability distributions of visual and textual variables extracted from the video data. In future work, it would be interesting to see how the methods scale up on a much larger dataset with more complex story structures than the educational cartoon videos for children.

Chapter 5

Story-aware Vision-Language Translation using Deep Concept Hiearachies

5.1 Overview

A language has been the most important way to communicate and store information for the past several thousands of years. Vision has been also used as a significant assistant method of linguistic representations for information delivery. The recent progress of information technology for the past two decades has caused the explosive increment of multimodal data such as video and images, and humans more frequently face multimodal data than simple text documents on the internet now a day. The task of automatically converting vision and language is considered as a key technique for valuable applications of these large-scale multimodal data (Kiros et al., 2014). However, a vision-language conversion with minimizing the distortion of semantic still remains a challenging issue due to the granularity difference between two modality representations and semantic gap (Turk, 2005).

Multimodal learning is a data-driven method for learning the relationships between two or more modalities and has been used as an approach suitable for the vision-language converting task. Many multimodal learning models have been reported as successful practical applications including image retrieval (Srivastava and Salakhutdinov, 2012; Blei and Jordan, 2003), automatic image annotation (Srivastava and Salakhutdinov, 2012; Blei and Jordan, 2003), and multimodal data classification (Karpathy et al., 2014). Latent dirichlet allocation (LDA) and its variants are a multimodal learning method based on a topic model for image retrieval and annotation (Nguyen et al., 2013; Blei and Jordan, 2003). Deep learning is a mainly used method for text and image association including deep belief networks (Socher et al., 2013), deep Boltzmann machines (Srivastava and Salakhutdinov, 2012), and convolution deep networks (Karpathy et al., 2014). Although these models showed successful applications, they mainly focused on enhancing the accuracy of the retrieved images and the annotated words rather than the constructing semantic knowledge at a higher level which can be used for further applications such as content summarization and knowledge organization. Furthermore, many conventional methods mostly concentrate on how efficiently the model is learned from a static and large-scale benchmarking dataset (Ordonez et al., 2011; Deng et al., 2009) but they seldom consider the change of the contents in the data with dynamic properties such as videos.

Here we view the vision-language converting task as a machine translation, i.e., a vision-language translation (V-L translation) and propose a method for V-L translation based on visually grounded knowledge constructed through learning the story contents from videos (Ha et al., 2014b). As a model for learning and organizing the knowledge from the video, we use a hierarchical model, i.e., a deep concept hierarchy (DCH), which characterizes the concepts and the concept relations contained in video stories as the knowledge. DCH consists of two kinds of layers; i) multiple conceptual layers and ii) a modality-dependent layer. The conceptual layers consist of one or more layers of concept variables for multiple conceptual levels and the variables in higher layers represent more abstract concepts. The modality-dependent layer contains the population of large number of microcodes encoding the higher-order relationships between two or more visual and textual variables. This model structure coincides with the grounded theory of human cognition system where a concept is grounded in the sensory-motor process including the modality-specific systems in which information is distributed to be stored (Quiroga, 2012). This structure enables DCH to represent the concepts with the probability distribution of the visual and the textual variables. As the model representation, DCH uses a flexible hypergraph structure.

The learning of DCH involves two main technical issues. One is to search a huge problem space represented with combinatorial features. Because a microcode store a higher-order association between two or more variables, the population and the connections between microcodes and concept variables is represented with a combinatorial feature space. The other is to deal with concept drift contained in the video data. A video story contains many concept relations, which change as the story unfolds. Furthermore, these concepts long-termly change over the progress of the stories. For handling these two issues, DCH uses a stochastic method based on a Monte-Carlo simulation for feasibly exploring the search space, i.e., a graph Monte-Carlo (graph MC) and incrementally learns the concepts and their change with a Bayesian update. Graph MC is a stochastic method to find optimal or suboptimal hypergraph structure by probabilistically adding and connecting the nodes using observed training data. The graph MC enables the model structure to flexibly grow and shrink for representing the concepts and this is a main difference from the other deep learning models. The weight of the constructed hypergraphs by graph MC are updated to reduce the Kullback-Leibler divergence between the observed data and the generated data by the model and this process is defined in terms of the Bayesian inference. Whenever observing new videos, DCH learns the concepts from the video by this mechanism and thus robustly tracing concept drift and continuously accumulating the knowledge on concepts. This learning is analogy for children to construct the visually grounded concepts from multimodal stimuli and to imitate the behaviors. The learned DCH provides visually grounded concept knowledge representation on the video story, a multimodal concept map.

This constructed knowledge is used to translate between the video scenes and the subtitles each other. While a given image is translated into words and sentences simply describing the objects of the image in conventional V-L translation, the proposed method translates the scenes into the subtitles considering the scene contents including the character relationships and situations. The knowledge on abstract concepts in DCH facilitates this contents-sensitive V-L translation. In specific, when scene images are given, the strength of the concepts in the scenes is estimated. Because each concept is represented with a mixture of the appearance distributions of words, word sets is generated to reflect the contents of the scenes. Finally, the subtitles corresponding to the scene are generated by a phrase-based approach. Text-to-vision translation is conducted by the same process except the image patch alignment. This translation is addressed in terms of Bayesian formulation in statistical machine translation. This process emulate a human crossmodal cognitive process to recall the subtitles when famous scenes is given after watching a movie, vice versa.

For evaluation, we used a famous cartoon video for children consisting of 183 episodes with 1,232 minutes of playing time, entitled ”Pororo.” Children’s cartoon videos are suitable for learning of DCHs since they have a simple and explicit story line and the relationships of the characters are not complex, compared to general videos. Moreover, the scene images are simple so that image preprocessing methods can show good performance. Experimental results show our method precisely translates the video scenes into the subtitles and vice versa. Furthermore, we investigate how the hierarchy of conceptual layers enhances the robustness of the model in concept drift. We present a novel representation of visually grounded knowledge for multimodal data, i.e., a multimodal concept map. Moreover, we demonstrate two applications: i) a visual story summarization of the video and ii) a generation of text story from arbitrarily composed scenes.

5.2 Vision-Language Conversion as a Machine Translation

5.2.1 Statistical Machine Translation

Machine translation is a computational linguistics including the task of translate text or speech from one language to another (Koehn, 2009). Machine translation has a very long history but statistical and data-driven approaches have been mainly used as the large-scale multilingual corpus become available. Recently, humans face diverse translation services such as google translation on the web or the mobile devices due to the progress of these statistical machine translation (SMT). The early methods for SMT were word-based models where words are used as atomic units. In word-based models such as IBM Candide project (Hutchins, 1993), the translation of sentences is considered as the task of word alignment. However, words are not suitable for the atomic units for translation because a word in a language can translate into two or more words in another language, or vice versa.

For resolving this issue, multiple words i.e., phrases have been used as the smallest unit for translation instead of one word, and many state-the-art SMT models belong to these phrase-based models (Durrani et al., 2013). The SMT can be mathematically defined using Bayes' rule:

$$p(\mathbf{e}|\mathbf{f}) = \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})} = p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \tag{5.1}$$

$$\mathbf{e}^* = \arg\max_e p(\mathbf{e}|\mathbf{f}) = \arg\max_e p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \tag{5.2}$$

where **e** and **f** denote the sentences in two languages, and $\mathbf{e}^*$ means the best sentence for **f**. In above equation, $p(\mathbf{f}|\mathbf{e})$ and $p(\mathbf{e})$ are the probability distributions of a translation model and a language model. For translation and language models, n-gram models and their variants have been widely used since proposed in early 1990s (Brown et al., 1992). In computational linguistics, an n-gram model is a probabilistic language model to predict the next word in a word sequence in the form of a (n-1)-order Markov model. Due to their simplicity and scalability, the n-gram models have been applied to diverse domain problems including computational biology and data compression as well as computational linguistics. However, n-gram models are difficult to avoid curse of dimensionality even if n is not large and thus they require many large corpus for precise prediction. In addition, various smoothing methods have been introduced for overcoming the problem of balancing the infrequent and the frequent grams by preventing the probability of infrequent grams from being zero (Mikolov et al., 2013a; Chen and Goodman, 1999). Recent studies on SMT have used a distributed representation of words and phrases for overcoming curse of dimensionality. Bengio et al. proposed the distributed representation of words and phrases using multilayer neural networks (Bengio et al., 2003) and it is reported that recurrent or deep neural networks have been successfully applied to learning language models and machine translation (Mikolov et al., 2010).

5.2.2 Vision-Language Translation

Vision-language translation (V-L translation) views the task of converting between visual and linguistic information as a type of machine translation. V-L translation involves many well-knownl problems including image annotation, photo description, and text-based image retrieval. While SMT methods learn the translation rules from bilingual corpus, it is required to learn the association rules between visual and linguistic data for V-L translation. In general, V-L translation is known to be more difficult than conventional translation between two languages due to the representation difference and semantic gap between two modalities (Fu et al., 2014). Similar to the SMT, V-L translation is defined with Bayes' rule:

$$p(\mathbf{w}|\mathbf{r}) = \frac{p(\mathbf{r}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{r})} = p(\mathbf{r}|\mathbf{w})p(\mathbf{w}) \tag{5.3}$$

$$p(\mathbf{r}|\mathbf{w}) = \frac{p(\mathbf{w}|\mathbf{r})p(\mathbf{r})}{p(\mathbf{w})} = p(\mathbf{w}|\mathbf{r})p(\mathbf{r}) \tag{5.4}$$

where r and w denote an image and a sentence. When introducing a concept model θ, an image and a sentence are translated from the corresponding sentences and images as follows:

$$w^* = \arg\max_{w} P(\mathbf{w}|\mathbf{r},\theta) = \arg\max_{w} P(\mathbf{r}|\mathbf{w},\theta)P(\mathbf{w}|\theta) \tag{5.5}$$

$$r^* = \arg\max_{r} P(\mathbf{r}|\mathbf{w},\theta) = \arg\max_{r} P(\mathbf{w}|\mathbf{r},\theta)P(\mathbf{r}|\theta) \tag{5.6}$$

where $\mathbf{w}^*$ and $\mathbf{r}^*$ denote the best sentence and image which describe the given image and sentence, respectively.

Early studies on V-L translation used topic models using the latent Dirichlet allocation (LDA) model. Blei et al. proposed an extension of the latent Dirichlet allocation (LDA) model for learning multimodal associations, i.e., correspondence-LDA (Corr-LDA) . In the Corr-LDA, the probabilities of both image and text variables are conditioned by the latent variable and thus multimodal associations are defined as topics for image annotation and retrieval(Blei and Jordan, 2003). Recent topic model-based approaches have focus on learning from multimodal data such as short video clips. Zhao et al. proposed a method for object detection from key frames of video clips using multimodal topic models (Zhao et al., 2013a). Fu et al. proposed the multimodal latent attribute topic model for transfer learning by learning semilatent attributes using video and audio modalities, and the model was applied to video classification (Fu et al., 2014). In addition, it has been reported that deep neural networks are successfully used for multimodal learning and V-L translation. Srivastva et al. proposed deep Boltzmann machines for learning large-scale multimodal data (Srivastava and Salakhutdinov, 2012). In this model, each modality is separately learned in the corresponding modality-specific layer. The modality-specific layers are integrated in higher layer encoding the associative information of two modalities. A zero-shot method based on deep learning models was proposed for crossmodal transfer learning by Socher et al. and was applied to image classification (Socher et al., 2013). Apart from these two approaches, various methods such as statistical methods (Li et al., 2008) and matrix factorization (Caicedo and González, 2012) have been used for learning the association between visual and linguistic modalities. However, these models are mainly applied to automatic image annotation, image segment labeling and text-based image retrieval but these applications were a little far from translation in a strict sense. More recent studies have handled the problems closer to the definition of translation. Kiros et al. proposed a multimodal log-bilinear model for generating sentences that describe the given images via image-text feature learning based on deep neural networks

(Kiros et al., 2014). However, most methods for V-L translation were applied to static large-scale annotated image database. Since they mainly use a fixed model structure, it is not easy to handle the translation in dynamic and increasing data such as videos. Some methods dealt with video data but they focused on classification or object detection rather than translation. In this chapter, we use large scale cartoon videos as the data and translate scene images and subtitles reflecting the video contents. This is more challenging because video data contain knowledge such as characters as well as concept changes included in the stories, i.e., concept drifts. In addition, the story content should be considered for more accurate translation. For handling these issues, it is required that the model for V-L translation uses more flexible representation which is suitable for incrementally learning from multimodal data streams.

5.3 Story-aware Vision-Language Translation using Deep Concept Hierarchies

5.3.1 Story-aware Vision-Language Translation

Tasks of converting between vision and language can be considered as a translation in the way that one representation is transformed into the other one with minimizing the distortion of its semantics. In this chapter, we view vision-language conversion as a translation, i.e., vision-language translation (V-L translation), which is formulated in terms of the statistical machine translation. In particular, we propose a method for translating between video scenes and subtitles considering the story contents which are represented as the concepts and call it story-aware V-L translation. This story-based translation is different from conventional methods for vision-language conversion in three aspects:

Sentences generated by conventional conversion

- Poby stands in front of the door. Trees are covered with snow.
- Loopy and Petty talk to Poby and Harry. Harry is on the head of Poby.
- There are four characters in the snow-covered field.

Sentences generated by the story-aware translation

- How are you today, Poby?
- Good morning. Loopy.
- Will you go to Pororo's house with us?

Figure 5.1: Example of sentences generated by conventional methods and story-aware vision-language translation.

i) Instead of sentences describing a scene image, the subtitles including situation explanations or dialogues are generated by the story-aware V-L translation for given scene images.

ii) The story-aware V-L translation synthesizes legible intermediate images by combining image patches representing the concepts associated with given sentences, which are used to retrieve similar original scene images.

iii) Given character information to the translation, the generated sentences and images varies depending on the characters despite the same query.

Figure 5.1 shows the difference between the sentences generated by the conventional V-L conversion and story-aware V-L translation. As shown in Figure 5.1, for example, ”How are you today, Poby” and ”Good morning, Loopy” are generated when Loopy and Poby are given as the observable concept variables, respectively.

Story-aware vision-language translation can be formulated from (5.3) and (5.4) by introducing a model parameter θ:

$$p(\mathbf{w}|\mathbf{r},\theta) = \frac{p(\mathbf{r}|\mathbf{w},\theta)p(\mathbf{w},\theta)}{p(\mathbf{r},\theta)} = \frac{p(\mathbf{r}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{r}|\theta)} \tag{5.7}$$

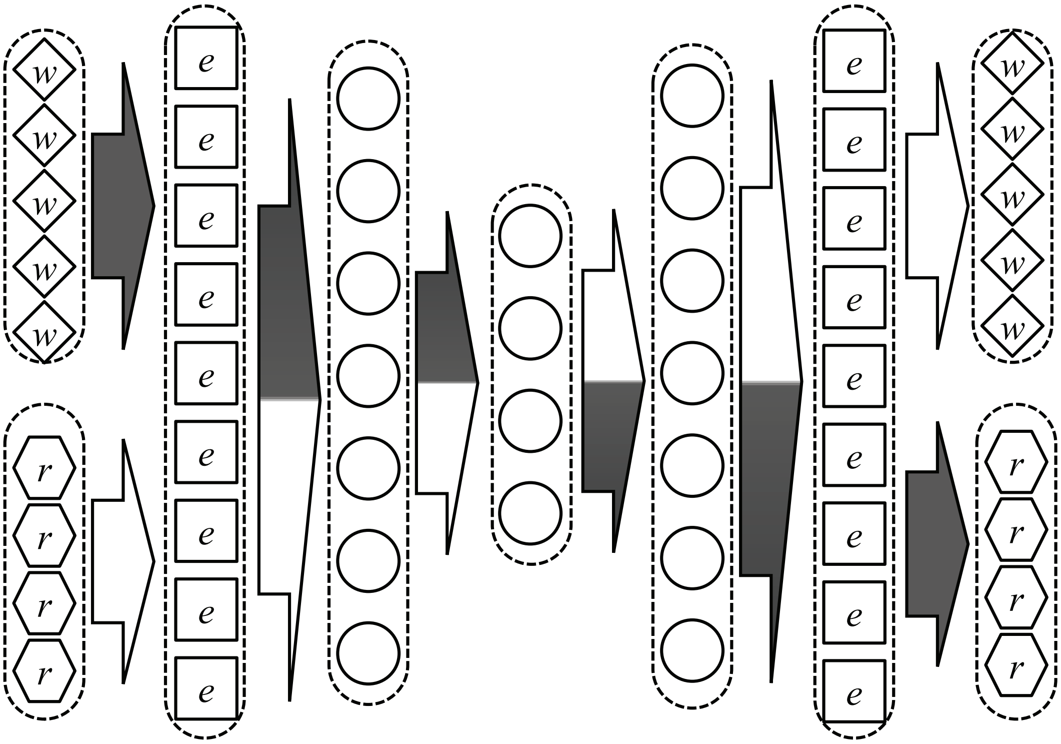


Figure 5.2: Vision-language translation by the crossmodal inference via a deep hypernetwork

$$p(\mathbf{r}|\mathbf{w},\theta) = \frac{p(\mathbf{w}|\mathbf{r},\theta)p(\mathbf{r},\theta)}{p(\mathbf{w},\theta)} = \frac{p(\mathbf{w}|\mathbf{r},\theta)p(\mathbf{r}|\theta)}{p(\mathbf{w}|\theta)} \tag{5.8}$$

Figure 5.2 presents vision-language crossmodal inference via deep hypernetworks. As shown in Figure 5.2, the mechanism of vision-language translation can be explained in terms of the encoding-decoding process. Scene-subtitles are encoded into multiple levels of visual-linguistic concepts and they are decoded into other modality representation from the concepts.

5.3.2 Vision-to-Language Translation

Vision-to-language translation denotes the generation of the subtitle sentences corresponding to given scenes as a query considering the video contents. Vision-to-language translation involves two steps: i) generating words associated with the scene image and ii) aligning the generated words.

The generation probability of the set of words associated with the given scene, in specific, is defined as the product of the probabilities of all the patches contained in the scene image for the word set by Bayes' rule:

$$P(\mathbf{w}|\mathbf{r},\theta) \propto \prod_{r_i \in \mathbf{r}} \{P(r_i|\mathbf{w},\theta)P(\mathbf{w}|\theta)\} \tag{5.9}$$

Ignoring $P(r|\theta)$ that are not independent on w and using log function for convenience, the best word set can be defined from (5.9):

$$\mathbf{w}{*} = \arg\max_{\mathbf{w}} \log P(\mathbf{w}|\mathbf{r},\theta) = \arg\max_{\mathbf{w}} \left\{ \sum_{r_i}^{|\mathbf{r}|} P(r_i|\mathbf{w},\theta) + |\mathbf{r}| \cdot \log P(\mathbf{w}|\theta) \right\} \tag{5.10}$$

Since we consider a model with two concept layers, θ is defined as the tuple of $\theta = (c^2, c^1, h)$. Using the phrase contained in each microcode as w, the first term can be computed as follows:

$$P(r_i|w,\theta) = \frac{\exp\left\{\sum_{h_m \in \mathbf{h}} \alpha_m \varphi(\mathbf{w}, \mathbf{c^1}, \mathbf{c^2}, h_m)\phi(r_i, \mathbf{c^1}, \mathbf{c^2}, h_m)\right\}}{\exp\left\{\sum_{h_m \in h} \alpha_m \varphi(\mathbf{w}, \mathbf{c^1}, \mathbf{c^2}, h_m)\right\}} \tag{5.11}$$

s.t

$$\varphi(\mathbf{w}, \mathbf{c^1}, \mathbf{c^2}, h_m) = \begin{cases} \frac{|w^{\mathrm{T}} \cdot h_m^w|}{|h_m^w|}, & if\ h_m \in \mathbf{h}^{\mathbf{c^1}} \text{ and } \omega(\mathbf{c^1}, \mathbf{c^2}) > \zeta \\ 0, & \text{otherwise} \end{cases}$$

and

$$\phi(r_i, \mathbf{c^1}, \mathbf{c^2}, h_m) = \begin{cases} 1, & if\ h_m \in h^{c^1},\ \omega(\mathbf{c^1}, \mathbf{c^2}) > \zeta, \text{ and } \min ED(r_i, h_m^r) < \xi \\ 0, & \text{otherwise} \end{cases}$$

where h^w_m and $\mathbf{h}^{\mathbf{c}^1}$ denote the words of the m-th microcode h_m and the microcode cluster associated with a $\mathbf{c}^1$ node, respectively. $ED(r_1, r_2)$ is the Euclidean distance between the vectors of two image patches. ζ and ξ are the thresholds of importance ratio of the character $\mathbf{c}^2$ in $\mathbf{r}$ and the patch distance. In this function, h is determined by $\mathbf{c}^1$ and $\mathbf{c}^2$. It means that the generated word set varies depending on the character despite the same scene. This definition allows the vision-to-language translation to be story-aware. By this process, the word set is generated including multiple words associated with the scene. The second term can be easily obtained by counting the frequency of the phrases in the model. The second step is aligning words. Since the generated word set mainly includes nouns and verbs, it is almost impossible to make sentences by aligning words in the set only. Therefore, we use the word set as a seed. Because each microcode encodes the consecutive words of subtitle sentences, next words are concatenated with the generated word one by one by the probability:

$$P(w_{n=i}|w_1, ..., w_{n-1}) = P(w_{n=i}|w_{n-k}, ..., w_{n-1}) = \frac{\exp\left\{\sum\limits_{h_m \in h} \alpha_m \varphi(w_{n-k:n-1} w_i, \mathbf{c}^1, \mathbf{c}^2, h_m)\right\}}{\exp\left\{\sum\limits_{h_m \in h} \alpha_m \varphi(w_{n-k:n-1}, \mathbf{c}^1, \mathbf{c}^2, h_m)\right\}} \tag{5.12}$$

where k is the length of Markov chain and we set k to 2. By introducing the concept layers into the indication function, we can generated character-specific dialogues and sentences. A dialogue sentence is generated by probabilistically concatenating words until a period is selected, and thus the length of the sentences is variable.

5.3.3 Language-to-Vision Translation

In contrast to vision-to-language translation, dialogues and sentences are transformed into intermediate images by the crossmodal inference, i.e. language-to-vision translation. Same as the vision-language translation, the translation consists of two steps including i) the selection of image patches associated with the given sentence and ii) the arrangement of the selected patches.

The generation probability of the patch set associated with given sentences is also defined as the product of the probabilities of generating words in the sentences for the patch set in the same way as the vision-to-language translation:

$$P(\mathbf{r}|\mathbf{w},\theta) \propto \prod_{w_i \in \mathbf{w}} \{P(w_i|\mathbf{r},\theta)P(\mathbf{r}|\theta)\} \tag{5.13}$$

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} \log P(\mathbf{r}|\mathbf{w},\theta) = \arg\max_{\mathbf{r}} \left\{ \sum_{w_i}^{|\mathbf{w}|} P(w_i|\mathbf{r},\theta) + |\mathbf{w}| \cdot \log P(\mathbf{r}|\theta) \right\} \tag{5.14}$$

The first term in (5.14) can be calculated in the same way:

$$P(w_i|\mathbf{r},\theta) = \frac{\exp\left\{ \sum_{h_m \in \mathbf{h}} \alpha_m \phi(\mathbf{r}, \mathbf{c}^1, \mathbf{c}^2, h_m) \varphi(r_i, \mathbf{c}^1, \mathbf{c}^2, h_m) \right\}}{\exp\left\{ \sum_{h_m \in \mathbf{h}} \alpha_m \phi(\mathbf{r}, \mathbf{c}^1, \mathbf{c}^2, h_m) \right\}} \tag{5.15}$$

In language-to-vision translation, we do not consider the alignment of the selected patches. Therefore, we select multiple patches based on the generation probability for all the patches in h, dissimilar to the vision-to-language translation. Selected patches are located with randomness to be an intermediate image.

5.4 Story-aware Vision-Language Translation on Catoon Videos

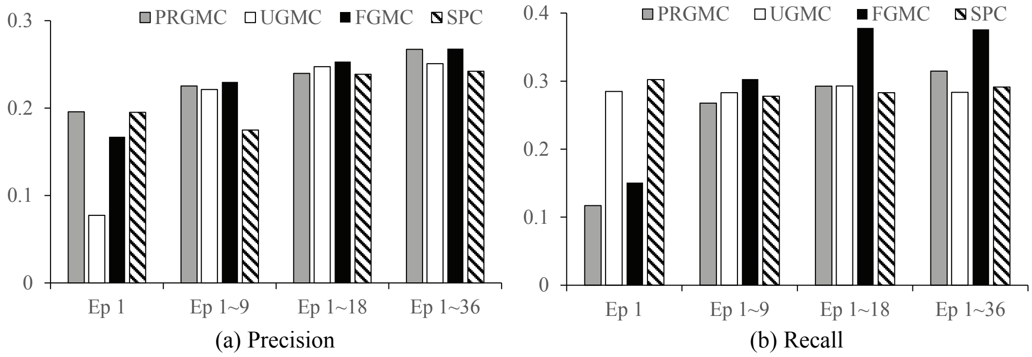
5.4.1 Data and Experimental Setting

We use two multimodal datasets in this section. One is *"Pororo"* dataset with 183 episodes, which is same to Chapter 4. By preprocessing, each scene is captured whenever a subtitle appears, transforming all the videos into the set of 16,000 utterance-scene pairs. A scene image is represented by a bag of image patches

Scenes	Episodes 1~13		Episodes 1~36	
	Original	And petty taught loopy how to ski.		
	UGMC	- did you ask me how to swim. - the end how grateful I think she is coming.	UGMC	- Wow petty that how that is not always so loopy taught if i can do fly it well. - How did you have to be that I could ski just.
	PRGMC	- end how was so happy - the end how did you I would	PRGMC	- How did you pororo you. - How about now you can I do not worry.
	FGMC	- To show how big you found - The end how grateful I am petty nice to lose careful	FGMC	- Harry realized that how that is it is dangerous - I thought that how that I could ski just
	SPC	- But how do someone stop. - The end how was it.	SPC	- How about now you can you give me that how that is great. - I will see let see how big.
	Original	Wow poby, you caught so many already.		
	UGMC	- Has been caught	UGMC	- Come out if you go in to hear you guys you have got a lot of fish I caught. - You have caught a lot today did you see you later.
	PRGMC	- Has been caught	PRGMC	- Everyone has caught a fish for dinner. - You have caught a lot today did you ask me how.
	FGMC	- What are you guys you have caught a lot. - What happened to ten everyone has caught a lot.	FGMC	- Poby caught a boat a secret that all the wind is so big. - You have caught a fish for the art diving.
	SPC	- Pororo no pororo has caught - She caught the first place	SPC	- You come with his new friend has caught a very interesting book recently - What about pororo has caught a lot of fish

Figure 5.3: Story-aware subtitle generation for given scene images as the increase of the observed videos. SPC denotes a model with no concept layer.

extracted by maximally stable external regions (MSER) (Matas et al., 2004) and each patch is defined as a feature vector using SIFT (Lowe, 2004) and RGB color. The other is a benchmarking dataset which is a database of tagged images from Flickr.com called MIR Flickr dataset (Huiskes and Lew, 2008). Among the dataset, we selected 10,000 images with category labels for evaluation. We used a DCH model with two concept layers. A microcode consists of two image patches and a phrase with three consecutive words. The image patches are selected by UGMC



(a) Precision

(b) Recall

Figure 5.4: Precision (a) and recall (b) of scene to sentence as the increase of the observed videos generation

and a phrase is selected with the maximum value of $P(v(x))$ of the words in the phrase.

5.4.2 Scene-to-Sentence Generation

Figure 5.3 shows the generated subtitle sentences for given scene images using concept knowledge on the video contents. The first image is an observed scene and the second is not observed by the models. For both the images, the model observing more videos generates not only more complex but also more descriptive sentences with more diverse words. Also, the model learned by FGMC provides more accurate and descriptive sentences, compared to those by PRGMC. We can obviously find this result from Figure 5.3, which presents the quantitative performance of story-aware scene to subtitle generation. We used the test set of 183 images by randomly selecting one image per episode image and the values are averaged after repeating 10 times of experiments. For evaluation, we used precision, recall, and F-score as

Table 5.1: Scene-to-sentence translation performance on the *Pororo* dataset

Score	cLDAs	mDBMs	PRGMC	UGMC	FGMC	SPC
Precision	0.020	0.101	0.267	0.251	0.268	0.242
Recall	0.240	0.152	0.315	0.284	0.376	0.291
F-score	0.037	0.121	0.289	0.266	0.313	0.264

measures. Three measures are defined as follows:

$$Precision = \frac{|C|}{|w'|}, Recall = \frac{|C|}{|w|}, \tag{5.16}$$

$$F - score = 2 \cdot \frac{Precesion \cdot Recall}{Precesion + Recall}, \tag{5.17}$$

where C denotes the set of correctly matched words. w and w' are an original and a generated subtitle sentence.

Table 5.1 shows the performances of translating scene images into sentences, compared to conventional multimodal models. For comparison, we used multimodal deep Boltzmann machines (mDBMs) (Srivastava and Salakhutdinov, 2012) and correpondence LDAs (cLDAs)(Xiao and Stibor, 2010). Because two models can generate textual tags only instead of sentences, the performances were measured using the generated tags. Dissimilar to mDBMs and corrLDAs, the performances of DHCs were computed using all the words in the generated sentences. The results in Table 5.1 are consistent with Figure 5.3. As shown in Table 5.1, we can find that DCHs not only provide subtitle sentences which cannot be generated by other two models but also show better performances than mDBMs and cLDAs. In specific, because two models mainly generate words frequently appearing in the observed

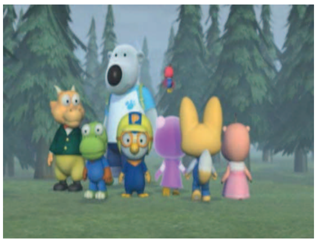
Scene images	Generated sentences (words)	
	Original	I am making a magic potion
	cLDA	is, I, you, crong, it,
	mDBM	cookie, cororong, uh, pipi, poyoyo
	DCH	- **i am making magic sorry i am making food** - **your magic wand already but i want to meet you sure**
	Original	Everyone headed into the forest
	cLDA	I, is, you, crong, the
	mDBM	cookie, crorong, uh, pipi, poyoyo
	DCH	**- you think everyone is better to the forest** **- we have to find everyone ran away**

Figure 5.5: Examples of sentences (words) generated by DCHs and other models

subtitles, including *it*, *is*, *a*, and *that*, they present low scores as shown in figure 5.5. Also, FGMC shows better performance than UGMC and PRGMC as the observed videos increase. This indicates that FGMC enables the concepts to be represented with more diverse and descriptive image patches and words. Also, we can indicate that the introduction of the hierarchy improves the model performance, comparing the results of DCHs to those of SPCs.

Table 5.2 presents the performance of image-to-text translation on the MIR Flickr.com dataset. We used FGMC as the method for learning DCHs. Dissimilar to Table 5.1, ovall performaces are lower than those of the Pororo dataset. We indicate that this is caused by the difficulty of representing natural images with

Table 5.2: Scene-to-sentence translation performance on Flickr.com dataset

Measure	cLDAs	mDBMs	DCHs(FGMC)	SPCs
ST(30)	0.241	0.213	0.198	0.263
ST(100)	0.533	0.341	0.412	0.579
Precision	0.004	0.005	0.011	0.004
Recall	0.101	0.090	0.192	0.111
F-score	0.007	0.009	0.019	0.006

visual features. Also, the tagged words are sparse compared to the subtitles of video scenes. Therefore, we confirmed that automatic tagging and sentence generation for images are still a very challenging problem. For effective comparison, we define a score, successful tagging (ST) that is 1 when the generated word set includes the real tag words of the given scenes, 0 otherwise. ST(n) denotes the average sucessful tagging when the size of a generate word set is n. Comparing the performances of each model, DCHs provide competitive results to cLDAs and outperform mDBMs. From Table 5.2, we indicate that the proposed model can provide competitive performances of vision-language translation on a natural images.

5.4.3 Sentence-to-Scene Generation

Figure 5.6 shows the generated intermediate images for given sentences as a query by story-aware vision-language translation. The generated scenes are synthesized by the weighted overlapping of image patches associated with the words in the sentences based on the constructed knowledge. This mechanism is inspired by the cross-modal reconstruction of mental imagery upon stimuli in human brains. When

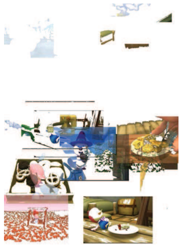
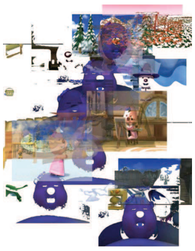
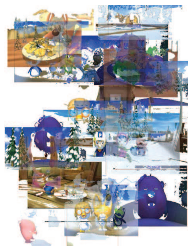
Query sentences	Episodes 1~52 (1 season)	Episodes 1~104 (2 seasons)	Episodes 1~183 (all seasons)
• Tongtong, please change this book using magic. • Kurikuri, Kurikuri-tongtong!			
• I like cookies. • It looks delicious • Thank you, loopy			

Figure 5.6: Generated intermediate images from given sentences as the increase of the observed videos

a child hears dialogue sentences, that is, he recalls the scenes or images related to the sentences. As the number of observed videos increase, the images become more complex and diverse. Comparing two query sentences, the first query sentences are related to *Tongtong*, a dragon magician and the second sentences are associated with *Loopy*, a chatter girl who likes cooking. Note that *Tongtong* does not appear until episode 56 and he casts "*Kurikuri*" for spell. Therefore, the images generated by the model learning from episode 1 to episode 52 seem to be unrelated to the first query sentences. However, once the concepts on *Tongtong* were constructed by observing *Tongtong*-related episodes, various images related to *Tongtong* recalled from the query sentences. Dissimilar to *Tongtong*, Loopy continuously comes on

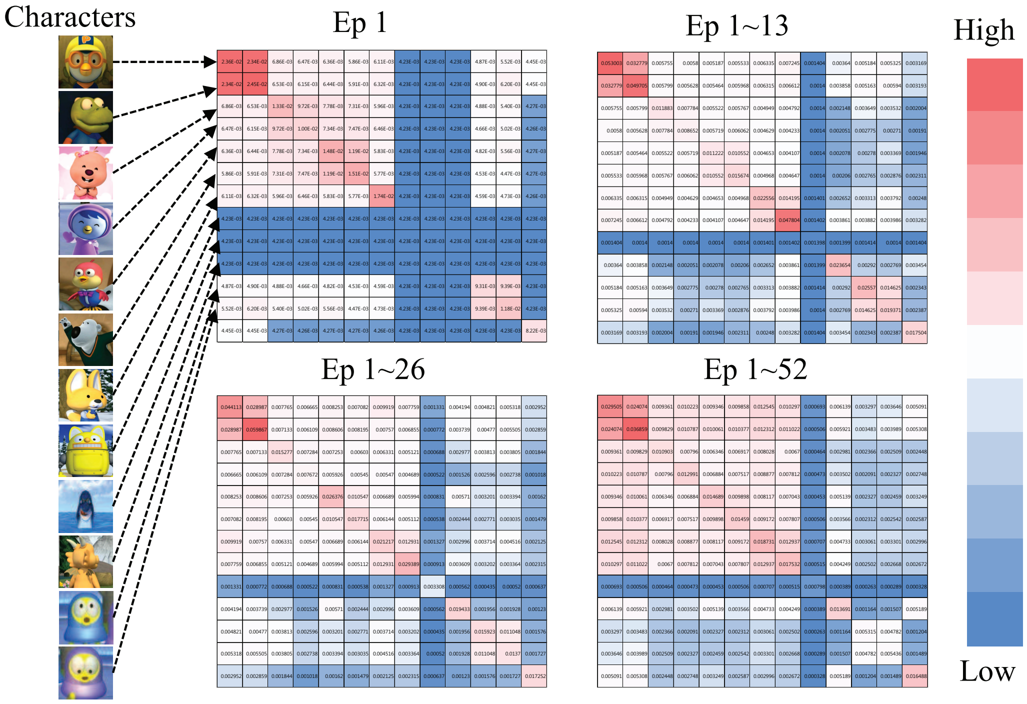


Figure 5.7: Changes of the relationships between main characters as the story proceeds

since episode 1 even if she less frequently appears than *Pororo* and *Crong*. In addition, she likes to make cookies. From the fact that the recalled images by the second query mainly mostly contain *Loopy*, cookies, and diverse objects related to *Loopy*, we indicate that the concept knowledge constructed by learning of DCHs enables the translation to be story-aware.

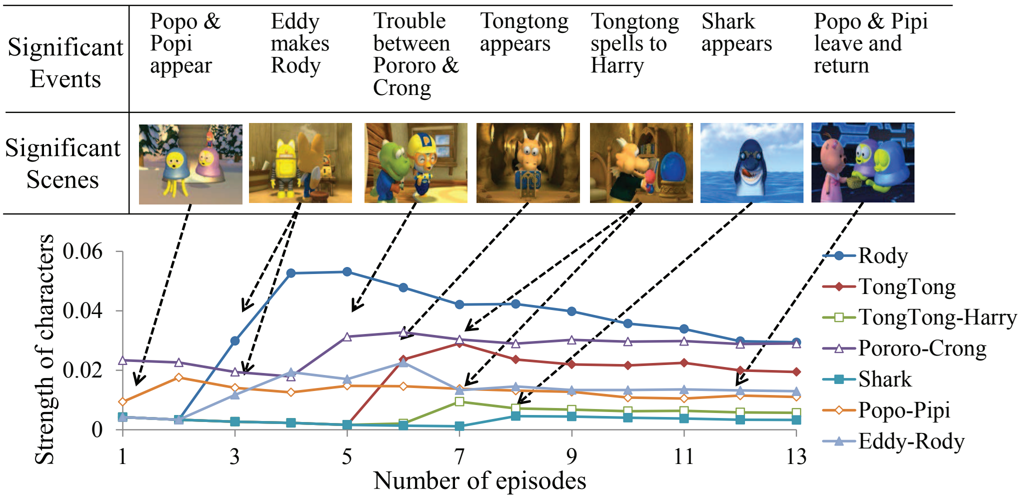


Figure 5.8: Visual-linguistic video summarization based on the changes of character relationships

5.4.4 Visual-Linguistic Story Summarization of Cartoon Videos

Figure 5.7 presents the change of the relationships between two characters in *Pororo* season 3 as the stories unfold. Through all the episodes, 13 main characters appear in the story and the character relationships continuously changes to determine the stories. In Figure Figure 5.7, red and blue denote strong and weak relationships, respectively. The relationships are computed using KL divergence between the distribution of the words and the patches associated with each character. As shown in Figure Figure 5.7, *Pororo* and *Crong* are strongly associated during all the episodes in general and this is consistent with the fact that *Pororo* is the older brother of *Crong* and they live together. In addition, The we can validate the fact that *Eddy* made *Rody* in the third episode and they live together after, comparing the result of one episode one to that of 13 episodes.

Figure 5.8 visualized the significant events corresponding to the relation changes as the stories proceed. This relation changes can be the summarization of the video story and the large changes mean the emergence of significant events.

5.5 Summary

We have presented a story-aware vision-language translation method based on the content knowledge, which is constructed via learning visual-linguistic concepts from cartoon videos. The story-aware translation is different from conventional vision-language conversion since it generates dialogues and narrations from scene images or immediate images from sentences, considering the contents of the observed videos. To achieve the story-aware vision-language translation, we have proposed a deep concept hierarchy (DCH) for learning the multimodal concepts used as knowledge for the translation. DCH represents grounded knowledge of vision and language by characterizing multiple levels of concepts with hypergraph structures. Unlike other deep learning models, the DCH structure can be flexibly and incrementally organized. This flexibility enables the model to handle concept drifts in stream data such as videos. To deal with the complexity problem for structure learning of DCH, we have proposed a graph Monte-Carlo method. In essence, the graph MC stochastically and constructively searches for a hierarchical hypergraph that matches the empirical distribution of the observed data. We compared three strategies for the graph MC: uniform graph Monte-Carlo (UGMC), poorer-richer graph Monte-Carlo (PRGMC), and fair graph Monte-Carlo (FGMC) depending on the probability of selecting vertices. Using the graph MC, DCH dynamically learns concepts from videos, thus automatically constructing knowledge with visual and linguistic representation on the video stories and is used to vision-language translation considering the video stories. Using the graph MC, DCH dynamically learns concepts from videos, thus automatically constructing knowledge with visual and linguistic representation on the video stories and is used to vision-language translation considering the video stories. Vision-language translation views the vision-language conversion as a machine translation and we formulated it in terms of the statistical machine translation. For allowing the translation to be story-aware, in addition, we introduced concept parameters into the formulation. This enables the translated results to vary despite the same queries, depending on characters as well as the amount of observed videos. We evaluated our method on cartoon videos, ”Pororo”, consisting of 183 episodes in addition to a benchmarking dataset.

Experimental results showed that DCH combined with the graph MC algorithm can represent and learn visual and linguistic concepts at multiple abstraction levels in the form of the probability distribution of visual-textual variables from the videos. Also, we confirmed that the hierarchy improved the concept representation. In addition, we found that PRGMC and FGMC work better in earlier and later steps of video observation by presenting the visual-language translation. Furthermore, we presented that the proposed DCH showed a competitive performance compared to conventional models including deep learning models.

The present work can be mainly extended to two directions. Overall, our experimental results demonstrate that DCH combined with the graph MC algorithms captures the mixed visual-linguistic concepts at multiple abstraction levels by sequentially estimating the probability distributions of visual and textual variables extracted from the video data. In future work, it would be interesting to see how the methods scale up on a much larger dataset with more complex story structures than the educational cartoon videos for children. The second direction is to be extended into learning from real-life sensor data stream for achieving human-level artificial intelligence based on lifelong learning. For achieving this, our method can be implemented using high-performance computing.

Chapter 6

Concluding Remarks

6.1 Summary of the Dissertation

We have proposed a multimodal hypernetwork with deep architecture for learning concept knowledge from dynamic multimodal data, i.e., deep hypernetwork, in this dissertation. A hypernetwork is a higher-order probabilistic graphical model using flexible hypergraph structures explicitly characterizing higher-order relationships among data variables.

In this dissertaion, hypernetworks are extend to represent associations among visual and linguistic features to model multimodal data such as annotated images and cartoon videos, called multimodal hypernetwork. By denoting an image patch and a word to a visual and a textual vertex, an hyperedge encodes the high-order relationship among visual-textual features, and thus multimodal hypernetwork represent the multimodal association contained in the observed data. We proposed a incremental method for learning from large-scale data and this is formulated in terms of the sequential Bayesian sampling. A multimodal hypernewtork was successfully applied to the text-to-image retrieval on 3000 images of the SBU photograph dataset.

Non-stationary multimodal data such as videos implicate concept drifts which are the changes of stories. In addition, a multimodal hypernetwork requires a large number of hyperedges for modeling large amount of data and this may cause the scalability problem. In this dissertation, for handling these two issues, we proposed a deep architecture of hypernetworks using a multiple layers of hypernetworks for learning from non-stationary multimodal data, called deep hypernetwork. Deep hypernetwork is different from conventional deep learning models with respect to the node connection between layers. The nodes between two adjacent layer are not fully connected in deep hypernetworks and this reduces the model complexity. Learning of deep hypernetwork involves two main technical issues; i) searching a huge combinatorial feature space representing a hypernetwork and ii) tracing concept drifts as the increase of the observed data. For efficient learning of hypernetworks, we proposed a stochastic method for constructing hypergraphs, i.e., graph Monte-Calro (graph MC). We defined the graph MC as three types such as uniform graph MC (UGMC), poorer-richer graph MC (PRGMC), and fair graph MC (FGMC) depending on its learning strategy. For dealing with concept drifts, the number of the nodes of each layer can flexibly changes as the learning proceeds. Also, the weights of the connections are updated while observing new data and this is formulated in terms of Bayesian inference. This learning mechanism allows a deep hypernetwork to effectivley model a hierarchy of concepts of the video contents and robustly trace the concept drift as the progress of the story.

Constructed concept hierarchies can be considered as the knowledge on the observed videos and this concept knowledge can be used for transforming between visual and linguistic contents. We view the vision-language conversion as the machine translation and formulated it in terms of the statistical machine translation, vision-language translation (V-L translation). Since the scenes and the subtitles are translated from each other considering the story, this makes the proposed method different from other vision-language conversion models, and we call it story-aware vision-langauge translation.

We used a famous cartoon video for children, Pororo, with 183 episodes as non-stationary multimodal data for evaluating the methods proposed in this dissertation. Experimental result showed that the deep hypernetworks outperform conventional multimodal hypernetwork with no hierarchy as well as conventional deep learning models. In contrast to topic models, the proposed model can precisely generated sentences from the given scene images while reflecting the observed story. Furthermore, our model can generated legible intermediate images from the given sentences, which can be used for retrieving original images. Also, we presented our method can robustly trace the concept drift implicated in the data by investigating the model structures, translating between the scenes and the subtitles, visualizing the development of visual-linguistic concept maps.

6.2 Directions for Further Research

The presented work is the start point for implementing lifelong learning to achieve human-level intelligence. For achieving this goal, this work should be extended into several directions. One is to add more modality such as audio into the model. The addition of audio enables the model to enhance the efficiency of language learning and to represent emotional features. For modeling human behaviors in real-world and real-life, it is essential for introducing more modality representing data generated diverse sensors equipped in smartphone and google glass into the model. The second direction is to improve theoretical soundness of the learning. The third is to introduce a dynamic architecture into the models by using prediction results as input information for learning. The fourth is to implement the method using high performance computing architecture for efficiently dealing with dynamic data. A hypernetwork uses the representation suitable for parallel and distributed computing since it consists of many hyperedges. The implementation based on high performance computing enables the model to efficiently learn from large-scale dynamic multimodal data.

Bibliography

Abbott, J. T., Austerweil, J. L., Griffiths, T. L., et al. (2012). Constructing a hypothesis space from the web for large-scale bayesian word learning. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Citeseer.

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, 20(1):303–330.

Austerweil, J. L. and Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, 120(4):817.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Berge, C. (1984). *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier.

Besson, J., Crawford, J., Parker, D., Ebmeier, K., Best, P., Gemmell, H., Sharp, P., and Smith, F. (1990). Multimodal imaging in alzheimer's disease. the relationship between mri, spect, cognitive and pathological changes. *The British Journal of Psychiatry*, 157(2):216–220.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892–895.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Caicedo, J. C., BenAbdallah, J., González, F. A., and Nasraoui, O. (2012). Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76(1):50–60.

Caicedo, J. C. and González, F. A. (2012). Multimodal fusion for image retrieval using matrix factorization. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 56. ACM.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2009 (CVPR 2009)*, pages 248–255. IEEE.

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. (2013). Can markov models over minimal translation units help phrase-based smt? In *ACL (2)*, pages 399–405.

Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Fan, J., Shen, Y., Yang, C., and Zhou, N. (2011). Structured max-margin learning for inter-related classifier training and multilabel image annotation. *IEEE Transactions on Image Processing*, 20(3):837–854.

Fearon, K., Arends, J., and Baracos, V. (2012). Understanding the mechanisms and treatment options in cancer cachexia. *Nature Reviews Clinical Oncology*, 10(2):90–99.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Feng, S., Manmatha, R., and Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages II–1002. IEEE.

Feng, Y. and Lapata, M. (2013). Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Fu, Y., Hospedales, T., Xiang, T., and Gong, S. (2014). Learning multi-modal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):303–316.

Gemmell, J., Bell, G., and Lueder, R. (2006). Mylifebits: a personal database for everything. *Communications of the ACM*, 49(1):88–95.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Guo, Z. and Wang, Z. J. (2013). An unsupervised hierarchical feature learning framework for one-shot image recognition. *IEEE Transactions on Multimedia*, 15(3):621–632.

Ha, J.-W., Eom, J.-H., Kim, S.-C., and Zhang, B.-T. (2007). Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis. In *Proceedings of the 9th Genetic and Evolutionary Computational Conference (GECCO 2007)*, pages 2709–2716. ACM.

Ha, J.-W., Jang, J. H., Kang, D.-H., Jung, W. H., Kwon, J. S., and Zhang, B.-T. (2009a). Gender classification with cortical thickness measurement from magnetic resonance imaging by using a feature selection method based on evolutionary hypernetworks. In *Proceedings of IEEE International Conference on Fuzzy Systems 2009 (FUZZ-IEEE 2009)*, pages 41–46. IEEE.

Ha, J.-W., Kim, B.-H., Kim, H.-W., C., Y. W., Eom, J.-H., and Zhang, B.-T. (2009b). Text-to-image cross-modal retrieval of magazine articles based on higher-order pattern recall by hypernetworks. In *Proceedings of International Symposium on Advanced Intelligence Systems 2009*, pages 274–277.

Ha, J.-W., Kim, B.-H., Lee, B., and Zhang, B.-T. (2010). Layered hypernetwork models for cross-modal associative text and image keyword generation in multimodal information retrieval. In *PRICAI 2010: Trends in Artificial Intelligence*, pages 76–87. Springer.

Ha, J.-W., Kim, K.-M., and Zhang, B.-T. (2014a). Automated construction of visual-linguistic knowledge via concept learning from cartoon videos. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence(AAAI 2015)(to appear)*.

Ha, J.-W., Kim, K.-M., and Zhang, B.-T. (2014b). Story-aware vision-language translation based on visually grounded knowledge learning. *Artificial Intelligence (in Preparation)*.

Ha, J.-W., Kim, S.-J., Kwon, J. S., and Zhang, B.-T. (2014c). Gradient-based learning applied to document recognition. *IEEE Transactions on Evolutionary Computation (in Revision)*.

Ha, J.-W., Lee, B.-J., and Zhang, B.-T. (2012). Text-to-image retrieval based on incremental association via multimodal hypernetworks. In *Proceedings of 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2012)*, pages 3245–3250. IEEE.

Halgren, E., Baudena, P., Heit, G., Clarke, M., and Marinkovic, K. (1994). Spatio-temporal stages in face and word processing. 1. depth recorded potentials in the human occipital and parietal lobes. *Journal of Physiology-Paris*, 88(1):1–50.

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5):5947.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM.

Hu, T., Xiong, H., Zhou, W., Sung, S. Y., and Luo, H. (2008). Hypergraph partitioning for document clustering: A unified clique perspective. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 871–872. ACM.

Huiskes, M. J. and Lew, M. S. (2008). The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 39–43. ACM.

Hutchins, J. (1993). Latest developments in machine translation technology beginning a new era in mt research. *MT Summit IV, Kobe Japan*.

Jeong, H., Yoo, Y., Yi, K. M., and Choi, J. Y. (2014). Two-stage online inference model for traffic pattern analysis and anomaly detection. *Machine Vision and Applications*, 25(6):1501–1517.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Jia, Y., Abbott, J. T., Austerweil, J., Griffiths, T., and Darrell, T. (2013). Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems*, pages 1842–1850.

Karimaghaloo, Z., Shah, M., Francis, S. J., Arnold, D. L., Collins, D. L., and Arbel, T. (2012). Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri using conditional random fields. *IEEE Transactions on Medical Imaging*, 31(6):1181–1194.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5.

Kiefer, M. and Barsalou, L. (2012). Grounding the human conceptual system in perception, action, and internal states. *Tutorials in Action Science*.

Kim, E.-S., Ha, J.-W., Jung, W. H., Jang, J. H., and Zhang, B.-T. (2011). Mutual information-based evolution of hypernetworks for brain data analysis. In *Proceedings of IEEE Congress on Evolutionary Computation 2011 (CEC 2011)*, pages 2611–2617. IEEE.

Kim, J.-K. and Zhang, B.-T. (2007). Evolving hypernetworks for pattern classification. In *Proceedings of IEEE Congress on Evolutionary Computation 2007 (CEC 2007)*, pages 1856–1862. IEEE.

Kim, S.-J., Ha, J.-W., Lee, B., and Zhang, B.-T. (2010). Evolutionary layered hypernetworks for identifying microrna-mrna regulatory modules. In *Proceedings of IEEE Congress on Evolutionary Computation 2010 (CEC 2010)*, pages 1–8. IEEE.

Kim, S.-J., Ha, J.-W., and Zhang, B.-T. (2013). Constructing higher-order mirna-mrna interaction networks in prostate cancer via hypergraph-based learning. *BMC Systems Biology*, 7(1):47.

Kim, S.-J., Ha, J.-W., and Zhang, B.-T. (2014). Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes. *Journal of Biomedical Informatics*.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In *Proceedings of of the 31st International Conference on Machine Learning (ICML 2014)*.

Klamt, S., Haus, U.-U., and Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Kok, S. and Domingos, P. (2009). Learning markov logic network structure via hypergraph lifting. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 505–512. ACM.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press.

Lauer, F., Suen, C. Y., and Bloch, G. (2007). A trainable feature extractor for handwritten digit recognition. *Pattern Recognition*, 40(6):1816–1824.

Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

Le Callet, P., Viard-Gaudin, C., and Barba, D. (2006). A convolutional neural network approach for objective video quality assessment. *IEEE Transactions on Neural Networks*, 17(5):1316–1327.

Le Roux, N. and Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Lee, B.-J., Ha, J.-W., Kim, K.-M., and Zhang, B.-T. (2013). Evolutionary concept learning from cartoon videos by multimodal hypernetworks. In *Proceedings of IEEE Congress on Evolutionary Computation 2013 (CEC 2013)*, pages 1186–1192. IEEE.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, 55(1):1–7.

Lehár, J., Krueger, A., Zimmermann, G., and Borisy, A. (2008). High-order combination effects and biological robustness. *Molecular Systems Biology*, 4(1).

Lewis, M. and Frank, M. C. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Li, H., Tang, J., Li, G., and Chua, T.-S. (2008). Word2image: towards visual interpreting of words. In *Proceedings of the 16th ACM International Conference on Multimedia*, pages 813–816. ACM.

Li, L.-J. and Fei-Fei, L. (2010). Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision*, 88(2):147–168.

Li, W., Duan, L., Xu, D., and Tsang, I. W.-H. (2011). Text-based image retrieval using progressive multi-instance learning. In *Proceedings of IEEE International Conference on Computer Vision 2011 (ICCV 2011)*, pages 2049–2055. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

Mei, T., Yang, B., Hua, X.-S., and Li, S. (2011). Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems (TOIS)*, 29(2):10.

Meltzoff, A. N. (1990). Towards a developmental cognitive science. *Annals of the New York academy of sciences*, 608(1):1–37.

Mesulam, M. (1994). Neurocognitive networks and selectively distributed processing. *Revue Neurologique*.

Mesulam, M.-M. (1998). From sensation to cognition. *Brain*, 121(6):1013–1052.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH 2010*, pages 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Muggleton, S. (2014). Alan turing and the development of artificial intelligence. *AI Communications*, 27(1):3–10.

Newman, M. and Barkema, G. (1999). Monte carlo methods in statistical physics chapter 1-4.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696.

Nguyen, C.-T., Zhan, D.-C., and Zhou, Z.-H. (2013). Multi-modal image annotation with multi-instance multi-label lda. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1558–1564. AAAI Press.

Nikitidis, S., Tefas, A., Nikolaidis, N., and Pitas, I. (2012). Subclass discriminant nonnegative matrix factorization for facial image analysis. *Pattern Recognition*, 45(12):4080–4091.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Paddock, S. M. and Savitsky, T. D. (2013). Bayesian hierarchical semiparametric modelling of longitudinal post-treatment outcomes from open enrolment therapy groups. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3):795–808.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM*

SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pages 159–168. ACM.

Phung, S. L. and Bouzerdoum, A. (2007). A pyramidal neural network for visual pattern recognition. *IEEE Transactions on Neural Networks*, 18(2):329–343.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597.

Rioux, D. and Van Meter, W. I. (1990). The abc's of awareness: A multimodal approach to relapse prevention. *Alcoholism Treatment Quarterly*, 7(3):77–89.

Roddick, J. F., Spiliopoulou, M., Lister, D., and Ceglar, A. (2008). Higher order mining. *ACM SIGKDD Explorations Newsletter*, 10(1):5–17.

Salakhutdinov, R. and Hinton, G. (2012). An efficient learning procedure for deep boltzmann machines. *Neural Computation*, 24(8):1967–2006.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943.

Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230.

Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Tan, H.-K., Ngo, C.-W., and Wu, X. (2008). Modeling video hyperlinks with hypergraph for web video reranking. In *Proceedings of the 16th ACM International Conference on Multimedia*, pages 659–662. ACM.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems 10*, pages 59–65.

Turk, M. (2005). Multimodal human-computer interaction. In *Real-time Vision for Human-Computer Interaction*, pages 269–283. Springer.

Xiao, H. and Stibor, T. (2010). Toward artificial synesthesia: Linking images and sounds via words. In *NIPS Workshop on Machine Learning for Next Generation Computer Vision Challenges*.

Yu, H., Chen, Y., Liu, J., and Jiang, X. (2014). Lifelong and fast transfer learning for gesture interaction. *Journal of Information and Computational Science*, 11(4):1023–1035.

Zaragoza, J. H., Sucar, L. E., Morales, E. F., Bielza, C., and Larranaga, P. (2011). Bayesian chain classifiers for multidimensional classification. In *Proceedings of 2011 International Joint Conference on Aritifical Intelligence (IJCAI 2011)*, volume 11, pages 2192–2197.

Zhang, B.-T. (2000). Bayesian evolutionary algorithms for learning and optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Workshop Program*, pages 220–222.

Zhang, B.-T. (2008). Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE Computational Intelligence Magazine*, 3(3):49–63.

Zhang, B.-T. (2013). Information-theoretic objective functions for lifelong learning. In *Proceedings of AAAI Spring Symposium: Lifelong Machine Learning*.

Zhang, B.-T., Ha, J.-W., and Kang, M. (2012). Sparse population code models of word learning in concept drift. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012)*, pages 1221–1226.

Zhang, B.-T. and Jang, H.-Y. (2005a). A bayesian algorithm for in vitro molecular evolution of pattern classifiers. In *DNA Computing*, pages 458–467. Springer.

Zhang, B.-T. and Jang, H.-Y. (2005b). Molecular programming: evolving genetic programs in a test tube. In *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation*, pages 1761–1768. ACM.

Zhang, B.-T., Ohm, P., and Mühlenbein, H. (1997). Evolutionary induction of sparse neural trees. *Evolutionary Computation*, 5(2):213–236.

Zhang, M.-L. and Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Zhang, R., Zhang, Z., Li, M., Ma, W.-Y., and Zhang, H.-J. (2005). A probabilistic semantic model for image annotation and multimodal image retrieval. In *Proceedings of the 10th IEEE International Conference on Computer Vision 2005 (ICCV 2005)*, volume 1, pages 846–851. IEEE.

Zhao, G., Yuan, J., and Hua, G. (2013a). Topical video object discovery from key frames by modeling word co-occurrence prior. In *Proceedings of 2013 IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR 2013), pages 1602–1609. IEEE.

Zhao, L., Shang, L., Gao, Y., and Jia, X. (2013b). Video behavior analysis using topic models and rough sets. *IEEE Computational Intelligence Magazine*, 8(1):56–67.

Zhou, D., Huang, J., and Schölkopf, B. (2006). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, pages 1601–1608.

Zimmerman, J. D. and Zeller, B. R. (1992). Imaginal, sensory, and cognitive experience inspontaneous recovery from alcoholism. *Psychological Reports*, 71(3):691–698.

초 록

최근 정보통신기술의 발달로 다양한 형태의 데이터가 급격히 증가중이며, 특히 정형화되고 단일 모달리티로 표현되는 과거의 데이터와는 달리, 최근 동적인 환경으로부터 생성되는 데이터들은 양은 물론, 멀티모달, 그리고 무정형성으로서 그 특징이 정의될 수 있다. 이러한 동적 멀티모달 데이터 학습 기술은 인공지능 분야에서 현재까지 해결할 수 없던 다양한 문제를 다루는 데 필수적이다. 그러나 여러 성공사례에도 불구하고 현존 기계학습기술은 이미지 검색 등과 같은 대량의 정적인 멀티모달 데이터로부터 표현되는 문제를 해결하는 데 집중해왔다.

하이퍼네트워크는 변수들간의 연관관계를 표현하는 하이퍼에지들의 집합으로 정의되는 하이퍼그래프 구조를 이용하여 데이터의 경험적 분포를 표현하는 모델이다. 그러나 하이퍼네트워크는 거대한 조합의 공간으로 정의되기 때문에, 대량의 멀티모달 데이터를 학습하기 위해서는 많은 수의 하이퍼에지들을 필요로 한다.

본 학위 논문에서는 복수의 하이퍼그래프들의 계층구조를 이용하여 비디오와 같은 동적 멀티모달 데이터를 효율적으로 학습하는 심층 모델인 심층 하이퍼네트워크 모델(deep hypernetwork)을 제안한다. 심층 하이퍼네트워크는 하이퍼그래프 계층 구조를 이용하여 다양한 수준의 추상화를 통해 대량 데이터의 학습에서 발생하는 기술적 문제들을 해결한다. 심층 하이퍼네트워크가 표현하는 거대한 문제공간을 효율적으로 탐색하기 위해 본 연구에서는 몬테카를로

시뮬레이션 기반의 확률적 그래프 구조 학습 기법인 그래프 몬테카를로(graph Monte Carlo) 기법을 제시한다. 고정된 모델구조를 기반으로 한 기존의 심층모델들과는 달리 심층 하이퍼네트워크는 모델 구조가 학습이 진행되는 동안 유동적으로 변화한다. 이러한 유연한 모델 구조는 동적 데이터에 포함된 개념이동(concept drift)를 다루기에 적합하다. 또한 학습을 통해 구축되는 개념은 멀티모달 데이터의 새로운 지식표현 방법으로 활용될 수 있으며 이러한 개념지식은 시각-언어간의 번역에 활용될 수 있다. 본 논문에서 우리는 시각-언어의 변환을 기계 번역으로 정의하고 학습된 개념지식을 이용하여 스토리 기반 시각-언어 번역기법을 제시한다.

제안하는 모델의 평가를 위해 본 학위논문에서는 벤치마킹 이미지 데이터베이스와 만화비디오를 이용하였다. 실험결과를 통해 우리는 제안하는 모델이 비디오의 컨텐츠를 효율적으로 다양한 수준에서의 시각-언어 개념으로 표현했음을 확인했다. 또한 계층 구조의 도입이 학습의 성능을 개선하는데 기여했음을 확인했다. 또한 학습된 정보의 정확성을 확인하기 위해 학습된 개념지식을 비디오 등장인물 분류와 비디오 장면과 자막간의 번역에 적용하였으며 실험결과를 통해 제안하는 모델이 기존 방법들에 비해 우수한 성능을 보였음을 확인하였다. 마지막으로 제안하는 모델의 평생학습을 위한 기술로서 갖는 의미와 인간수준 지능기술 구현을 위한 방향을 논의하였다.

주요어: 심층 하이퍼네트워크, 고차 그래프 모델, 동적 멀티모달 데이터, 멀티모달 개념 학습, 확률적 하이퍼그래프 생성, 점진적 학습, 시각-언어 번역

학 번: 2006-21317