Layered Hypernetwork Models for Cross-Modal Associative Text and Image Keyword Generation in Multimodal Information Retrieval

Jung-Woo Ha, Byoung-Hee Kim, Bado Lee, and Byoung-Tak Zhang

Biointelligence Lab, School of Computer Science and Engineering,
Seoul National University,
599 Gwanak-ro, Gwank-gu, Seoul 151-744, Korea
{jwha,bhkim,bdlee,btzhang}@bi.snu.ac.kr

Abstract. Conventional methods for multimodal data retrieval use text-tag based or cross-modal approaches such as tag-image co-occurrence and canonical correlation analysis. Since there are differences of granularity in text and image features, however, approaches based on lower-order relationship between modalities may have limitations. Here, we propose a novel text and image keyword generation method by cross-modal associative learning and inference with multimodal queries. We use a modified hypernetwork model, i.e. layered hypernetworks (LHNs) which consists of the first (lower) layer and the second (upper) layer which has more than two modality-dependent hypernetworks and one modality-integrating hypernetwork, respectively. LHNs learn higher-order associative relationships between text and image modalities by training on an example set. After training, LHNs are used to extend multimodal queries by generating text and image keywords via cross-modal inference, i.e. text-to-image and image-to-text. The LHNs are evaluated on Korean magazine articles with images on women fashions and life-style. Experimental results show that the proposed method generates vision-language cross-modal keywords with high accuracy. The results also show that multimodal queries improve the accuracy of keyword generation compared with uni-modal ones.

Keywords: hypernetwork, layered hypernetwork, cross-modal generation, vision-language, text-to-image, image-to-text, multimodal information retrieval.

1 Introduction

Recently, cross-modal learning methods have been considered as a major approach for multimodal information retrieval such as video, image, and article retrieval as well as automatic tagging and annotation [1-3]. Because there are differences of granularity in text and image features, however, simple approaches based on text-image relations have the limitation to learn. As a model to learn higher-order cross-modal associations, we used hypernetwork models in the previous study [4]. A hypernetwork is a higher-order probabilistic graphical model which has properties including glocality, compositionality, self-assembly, and recall-memory [5]. In the previous study, we

B.-T. Zhang and M.A. Orgun (Eds.): PRICAI 2010, LNAI 6230, pp. 76–87, 2010.

showed that images could be retrieved with multimodal queries by text-to-image inference with trained hypernetworks [4].

In this study, we propose a novel modified hypernetwork model, layered hypernetworks (LHNs), which conducts cross-modal associative learning and inference including image-to-text as well as text-to-image for multimodal information retrieval. An LHN is a hypernetwork model with a hierarchical structure of two layers of hypernetwork. While the first layer is composed of modality-dependent hypernetworks, only one hypernetwork exists in the second layer which represents relationships between the text modality and the image modality. The hierarchical structure make LHNs analyzed with efficiency compared with conventional hypernetworks. Trained LHNs can generate both text and image keywords by cross-modal associative inference with multimodal queries. In addition, generated visual and textual keywords are used to retrieve articles by comparing them with text terms in document and visual words in images of articles. We use 983 Korean magazine articles with 8,763 images on women fashion and life-style as multimodal data. In this study, our contributions are summarized as follows.

1. We propose a novel modified hypernetwork named to layered hypernetwork for cross-modal associative learning and inference.
2. We propose a method to generate visual and textual keywords based on text-to-image and image-to-text cross-modal association.
3. We apply the proposed model to magazine article retrieval.

The rest of this paper is organized as follows. In Section 2, we summarize related works. Also, we explain layered hypernetworks for cross-modal association in Section 3 and propose a method for cross-modal keyword generation in Section 4. Section 5 presents the experimental results. Finally, we present concluding remarks in Section 6.

2 Related Works

As multi-media data increase explosively, multimedia data retrieval has been important problem in information retrieval such as video, image and articles. As an approach, cross-modal associative learning has been applied to multimodal data retrieval although cross-modal learning is from cognitive science and neuroscience [6]. Snoek *et al.* proposed concept-based video retrieval method [7] and Yan *et al.* studied a multimodal retrieval approach including text and image for broadcast new video [8]. D. Li *et al.* [9] suggested cross-modal association based factor analysis method as alternatives to Latent Semantic Indexing (LSI) and Canonical Correlation Analysis (CCA). Ferecatu *et al.* showed that the joint use of visual features and concept-based features with relevance feedback scheme improves the quality of the cross-modal image retrieval [10]. Goh *et al.* proposed an image retrieval method based on multimodal concept-dependent active learning [2]. Also, auto-annotation on unlabeled images and objects in images is carried out by using hierarchical latent Dirichlet allocation model [11]. In addition, human-computer interaction (HCI) is a research where cross-modal learning is considered as an essential element. In HCI, various modalities are studied including speeches and gestures. Quek *et al.* studied multimodal human discourse in aspect of gesture and speech [12]. Christoudias *et al.* proposed co-training method of multimodal data to construct multimodal interface [13]. However, conventional studies on cross-modal learning are usually based on lower-order co-occurrence on modalities rather than higher-order relations. Therefore, we propose a cross-modal learning method based on higher-order inter-modal relationships in this paper.

3 Cross-Modal Associative Learning Models

3.1 Hypernetwork Model

A hypernetwork is a bio-inspired probabilistic graphical model based on hypergraph models. The properties of the hypernetwork model are summarized as three aspects: glocality, compositionality and self association based on randomness and recall [5].

1. Glocality: A hypernetwork consists of hyperedges with various orders. Lower-order hyperedges can represent general information and higher-order ones include more specific and local information.
2. Compositionality: A hypernetwork represents a huge structured combinatorial space. By learning based evolutionary strategy, a hypernetwork explores the combinatorial problem space.
3. Self association: The structure of hypernetworks is self-organized by evolutionary computation based on random selection. Self association makes the hypernetwork act like a recall memory.

Formally, a hypernetwork H is defined as $H = (V, E, W)$ where V, E, and W are a set of vertices, hyperedges, and weights. In hypernetworks, a vertex means a value of attributes and a hyperedge represents the combination of more than two vertices with its own weight. The number of vertices in a hyperedge is called cardinality or order of a hyperedge and k-hyperedge denotes a hyperedge with k vertices. When orders of all hyperedges are k, we call it k-hypernetwork. Therefore hypernetworks can represent higher-order relationships among large numbers of attributes.

Since a hypernetwork can be regarded as a probabilistic associative memory model to store segments of a given data set $D=\{\mathbf{x}^{(n)}\}_{n=1}^{N}$ i.e. $\mathbf{x}=\{x_1,x_2,\ldots x_m\}$, a learned hypernetwork can retrieve a data sample later. When $I(\mathbf{x}^{(\mathrm{n})}, E_i)$ denotes a function which yields the combination or concatenation of elements of E_i as (2), then, the energy of hypernetwork is defined as follows:

$$\varepsilon(\mathbf{x}^{(n)};W) = -\sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(n)}, E_i)\ , \tag{1}$$

$$I(\mathrm{x}^{(n)}, E_i) = x_{i1}^{(n)} x_{i2}^{(n)} \ldots x_{ik}^{(n)}, \tag{2}$$

where $w_i^{(k)}$ is a weight of i-th hyperedge E_i with k-order, $\mathbf{x}^{(n)}$ means the n-th stored pattern of data and E_i is $\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$. Then, the probability of the data generated by a hypernetwork $P(D|W)$ is given as a Gibbs distribution:

$$P(D \mid W) = \prod_{n=1}^{N} P(\mathbf{x}^{(n)} \mid W)\,, \tag{3}$$

$$P(\mathbf{x}^{(n)} \mid W) = \frac{1}{Z(W)} \exp\left(-\varepsilon(\mathbf{x}^{(n)}; W)\right), \tag{4}$$

where $Z(W)$ is a partition function. In addition, the partition function $Z(W)$ is formulated as follow:

$$Z(W) = \sum_{\mathbf{x}^{(m)} \subset D} \exp\left\{ \sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(m)}, E_i) \right\}. \tag{5}$$

That is, a hypernetwork is represented with a probability distribution of combination of variables with weights as parameters when we consider attributes in data as random variables. Considering that learning of hypernetworks is selecting hyperedges with high weight value, the learning can be considered as the process for maximizing log-likelihood. Leaning from data is regarded as maximizing probability of weight parameter of a hypernetwork for given data. Given data, probability of a weight set of hyperedges $P(W|D)$ is defined as follows:

$$P(W \mid D) = \frac{P(D \mid W) P(W)}{P(D)}. \tag{6}$$

According to (4) and (6), then, likelihood is defined as

$$\prod_{n=1}^{N} P(\mathbf{x}^{(n)} \mid W) P(W) = \left(\frac{P(W)}{Z(W)} \right)^{N} \exp\left\{ -\sum_{n=1}^{N} \varepsilon(\mathbf{x}^{(n)} \mid W) \right\}. \tag{7}$$

Ignoring $P(W)$, maximizing the argument of exponential function is obtaining maximum likelihood. Using log function,

$$\arg\max_{W} \left[\log \left\{ \prod_{n=1}^{N} P(\mathbf{x}^{(n)} \mid W) \right\} \right] = \arg\max_{W} \left\{ \sum_{n=1}^{N} \sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(n)}, E_i) - N \log Z(W) \right\}. \tag{8}$$

More explanations on the derivative of the log-likelihood are showed in [5]. Therefore, log-likelihood of hypernetwork can be maximized by decreasing the difference of hyperedges from a given data set.

3.2 Layered Hypernetworks

An LHN is a hypernetwork with hierarchical structures and the model consists of two layers. The first layer is a modality layer and the second one is an integrating layer. When data consisting of more than one modality are given, the attributes of given data are partitioned based on modalities. Hypernetworks in the first layer are built by sampling from attributes of each modality and the number of hypernetwork in the first layer is equal to the number of modalities. Dissimilar to the first layer, only one hypernetwork exists in the second layer. The second layer hypernetwork is built by combining hyperedges randomly selected from modality-dependent hypernetworks in the first layer. Therefore the hypernetwork in the second layer represents the relationship between several modalities. Same as conventional hypernetworks, formally, the second-layer hypernetwork is defined with the energy function when a weight vector is given as a parameter. When given a data set *D* consisting of two modalities, $D=\{\mathbf{x}^{(n)}\}_{n=1}^{N}=\{(\mathbf{m}^1,\mathbf{m}^2)^{(n)}\}_{n=1}^{N}$, the energy of the second-layer hypernetwork $\varepsilon(\mathrm{x}^{(n)};W)$ generated from *k*-hypernetworks in the first-layer is defined as follows:

$$\varepsilon(\mathbf{x}^{(n)};W)=\varepsilon\left\{(\mathbf{m}^1,\mathbf{m}^2)^{(n)};W\right\}=-\sum_{i=1}^{|E|}w_i^{(k)}I\left\{(\mathbf{m}^1,\mathbf{m}^2)^{(n)},E_i\right\}, \tag{9}$$

where $\mathbf{m}^1$ and $\mathbf{m}^2$ are vectors of each modality variable which constitute the *n*-th data sample $\mathrm{x}^{(n)}$. Same as (4), then, the probability of generating *n*-th data with two modalities, $P(\mathbf{x}^{(n)}|W)$ is defined as follows:

$$P(\mathbf{x}^{(n)}\,|\,W)=\frac{1}{Z(W)}\exp\left[-\varepsilon\left\{(\mathbf{m}^1,\mathbf{m}^2)^{(n)};W\right\}\right]. \tag{10}$$

Assuming that $\mathbf{m}^1$, $\mathbf{m}^2$ are text and image modality respectively, similar as conventional hypernetwork, the probability of data generated by layered hypernetworks, $P(D|W)$ is defined as follows:

$$\begin{aligned}P(D\,|\,W)=P(T,I\,|\,W)&=P(T\,|\,I,W)P(I\,|\,W)\\&=P(I\,|\,T,W)P(T\,|\,W).\end{aligned} \tag{11}$$

Formula (11) means that cross-modal inferences between text and image are carried out by learning parameters of hypernetworks. Figure 1 shows the architecture of LHNs.

3.3 Cross-Modal Associative Learning of Layered Hypernetworks

3.3.1 Learning of the First-Layer Hypernetworks

Learning of the first-layer hypernetworks is similar to the learning of conventional hypernetworks [4-5] except building a hypernetwork per one modality. At first, multimodal data are separated by modalities. In this study, an article data with unique id are divided into vectors of TF-IDF values from documents and vectors of histogram value from included images. The unique id is used to combine hyperedges of each modality in learning of the second-layer hypernetwork. Building a hypernetwork is carried out by generating hyperedges from each modality and hyperedges are generated by selecting and combining the attributes with non-negative values with randomness for each modality. The reason to select the attributes with non-negative values is that hyperedges where values of all vertices are zero may be generated with high probability because most attributes have zero value due to sparsity of data. As explained in Section 3, learning of hypernetwork is sampling hyperedges which are less different from data set. Details of building and learning a hypernetwork are explained in [5]. As learning continues, the structure of a hypernetwork fits the distribution of given data more. The constitution of hyperedges, the structure of a hypernetwork, is determined by their weights which reveal the fitness with training data set. In this study, we define the weight of a hyperedge, w, as follows:

$$w = \frac{C}{\#\text{of matched training samples} + k}, \tag{12}$$

where k denotes order of a hyperedge and C is an arbitrary constant. According to (12), hyperedges with unique information get higher weights by definition. Also, hyperedges with low weight values are eliminated and the erased amounts of hyperedges are regenerated from training set.

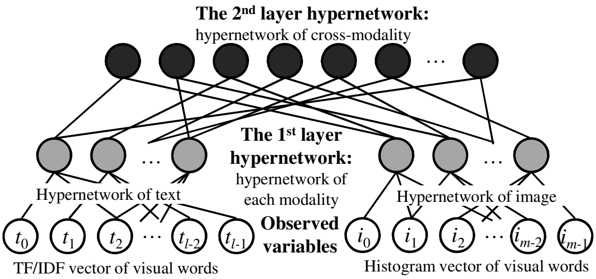
The 2nd layer hypernetwork:
hypernetwork of cross-modality
The 1st layer hypernetwork:
hypernetwork of each modality
Hypernetwork of text
Hypernetwork of image
Observed variables
t_0 t_1 t_2 … t_{l-2} t_{l-1}
i_0 i_1 i_2 … i_{m-2} i_{m-1}
TF/IDF vector of visual words
Histogram vector of visual words

Fig. 1. Architecture of layered hypernetwork models

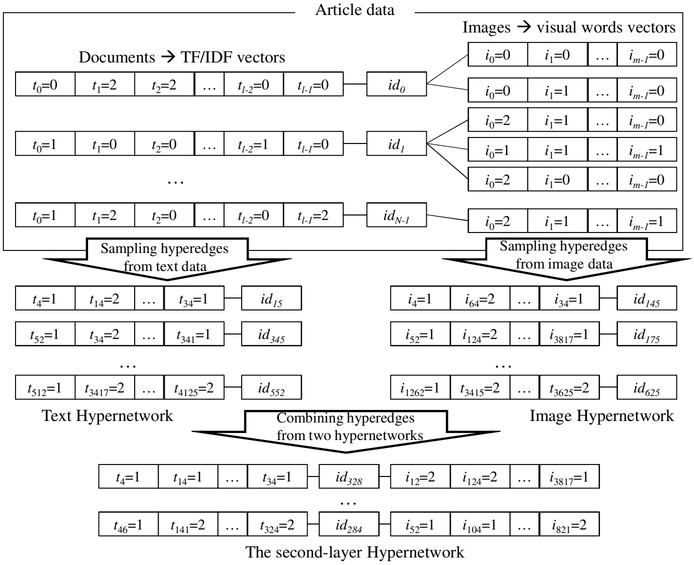
Article data
Images → visual words vectors
Documents → TF/IDF vectors
$t_0=0$ | $t_1=2$ | $t_2=2$ | … | $t_{l-2}=0$ | $t_{l-1}=0$ | id_0
$i_0=0$ | $i_1=0$ | … | $i_{m-1}=0$
$i_0=0$ | $i_1=1$ | … | $i_{m-1}=0$
$i_0=2$ | $i_1=1$ | … | $i_{m-1}=0$
$t_0=1$ | $t_1=0$ | $t_2=0$ | … | $t_{l-2}=1$ | $t_{l-1}=0$ | id_1
$i_0=1$ | $i_1=1$ | … | $i_{m-1}=1$
$i_0=2$ | $i_1=0$ | … | $i_{m-1}=0$
…
$t_0=1$ | $t_1=2$ | $t_2=0$ | … | $t_{l-2}=0$ | $t_{l-1}=2$ | id_{N-1}
$i_0=2$ | $i_1=1$ | … | $i_{m-1}=1$
Sampling hyperedges from text data
Sampling hyperedges from image data
$t_4=1$ | $t_{14}=2$ | … | $t_{34}=1$ | id_{15}
$t_{52}=1$ | $t_{34}=2$ | … | $t_{341}=1$ | id_{345}
…
$t_{512}=1$ | $t_{3417}=2$ | … | $t_{4125}=2$ | id_{552}
$i_4=1$ | $i_{64}=2$ | … | $i_{34}=1$ | id_{145}
$i_{52}=1$ | $i_{124}=2$ | … | $i_{3817}=1$ | id_{175}
…
$i_{1262}=1$ | $t_{3415}=2$ | … | $t_{3625}=2$ | id_{625}
Text Hypernetwork
Image Hypernetwork
Combining hyperedges from two hypernetworks
$t_4=1$ | $t_{14}=1$ | … | $t_{34}=1$ | id_{328} | $i_{12}=2$ | $i_{124}=2$ | … | $i_{3817}=1$
…
$t_{46}=1$ | $t_{141}=2$ | … | $t_{324}=2$ | id_{284} | $i_{52}=1$ | $i_{104}=1$ | … | $i_{821}=2$
The second-layer Hypernetwork

Fig. 2. The process of making and learning a layered hypernetwork

3.3.2 Learning of the Second-Layer Hypernetwork

Learning of the second-layer hypernetwork is to generate hyperedges which represent high-order relationships between modalities from the first-layer hypernetworks. Hyperedges of the second-layer hypernetwork are generated by combining hyperedges of hypernetworks in the first-layer. In combining, hyperedges from different modalities with the same id are merged into a new hyperedge. The weight of the generated hyperedge is obtained by comparing with training set same and hyperedges with low weights are also eliminated from the hypernetwork same as learning in the first layer. Then, the generated hypernetwork is evaluated with training data set. Figure 2 shows the process of making and learning a layered hypernetwork. In addition, algorithm of building and learning the second-layer hypernetwork is presented in detail in Figure 3. In our method, learning process finishes after fixed number of epochs.

```
$H_T$: hypernetwork from text data, $H_I$: hypernetwork from image data,
$H_L$: layered hypernetwork   $R$: replacing rate of hyperedges with low weights
CR: combining rate of hyperedges of $H_I$ with a hyperedge of $H_T$
$H_T \leftarrow$ makeHypernetwork($T$); $H_I \leftarrow$ makeHypernetwork($I$)
For  $i \leftarrow 1$ until end condition
   $H_T \leftarrow$ learningHypernetwork($T$); $H_I \leftarrow$ learningHypernetwork($I$);
   $H_T \leftarrow$ removeLowedges($R$); $H_I \leftarrow$ removeLowedges($R$); $H_L \leftarrow \{\}$;
   For $j \leftarrow 1$ to $| H_T |$
        $E_T \leftarrow$ the $j$-th hyperedge of $H_T$
        For k $\leftarrow 1$ to CR
             $E_I \leftarrow$ a randomly selected hyperedge with same id to $E_T$ from $H_I$ ;
                 $E_L \leftarrow E_T \cup E_I$; $H_L \leftarrow H_L \cup E_L$
        End For
   End For
   $H_L \leftarrow$ removeLowedges($R$); $H_L \leftarrow$ learningHypernetwork($T$, $I$);
   evaluate($H_L$, $I$, $T$)
   $H_T$ = Resampling($T$, $R$); $H_I$ = Resampling($I$, $R$)
End For
```

Fig. 3. Algorithm of building and learning a layered hypernetwork. Details of functions for learning are explained in our previous studies [4-5].

4 Cross-Modal Inference for Image and Text Keyword Generation

Trained LHNs can generate both text terms and visual words with given multimodal queries by cross-modal associative inference. Cross-modal associative generation is divided into two types such as text-to-image to generate a set of visual words for given text terms and image-to-text generation to reconstruct a set of text terms with visual words. In image-to-text, the generated set of text terms is composed of text terms in hyperedges of the second-layer hypernetwork whose vertices include at least one visual word in the given set of visual words. To select text terms, we define a score based on co-occurrence of text terms and visual words. When a visual word set Q, the score $s_{Idx(i),En}$ of the i-th text term in the n-th hyperedge E_n of the second-layer hypernetwork is defined as follow:

$$s_{Idx(i),E_n} \begin{cases} \dfrac{x_{Idx(i)}{}^2 \times w_n}{|Q - E_n| \times C + 1} & (Q \cap E_n \neq \varnothing) \\ 0 & (Q \cap E_n = \varnothing) \end{cases}, \tag{13}$$

where $x_{Idx(i)}$ is the value of text term attribute whose index is $Idx(i)$, $Idx(i)$ denotes the index in the vector representation of the i-th text term of a hyperedge E_n, w_n means weight of E_n, $|Q$ - $E_n|$ is the size of the relative complement, and C is a arbitrary constant for penalty. Therefore, $s_{Idx(i)}$ is obtained by summing for all hyperedges as follow:

$$s_{Idx(i)} = \sum_{n=1}^{|E|} s_{Idx(i),E_n}, \tag{14}$$

where $|E|$ denotes the number of hyperedges in the second-layer hypernetwork. According to (13), as a hyperedge includes more visual words in given visual word set, the score of text terms in the hyperedge gets larger. Then, text terms with higher score are included candidates for generated text keywords.

Same as image-to-text, a set of visual words are generated with trained layered hypernetwork and given text terms.

5 Experimental Results

5.1 Data and Experimental Setups

We use 983 articles with 8,673 images from three Korean magazines on female fashion name to 'luxury', 'beauty life' and 'haute' respectively as training data from a company named to ddh co. As preprocessing for modeling, documents in articles are converted to vectors of TF-IDF values of 5,000 text terms which are selected by

Table 1. The parameters used for the experiment

Parameters	**Value**
Order (text, image)	(20, 20)
Replacing rate	0.1
Sampling rate (text, image)	(20, 10)
Combining rate	10
Num. of iteration	5

Combining rate means the combining number of hyperedges of one modality hypernetwork for a hyperedge of the other modality hypernetwork in learning of the second layer. Sampling rate denotes the size of sampled hyperedges from a training data sample. Replacing rate is eliminated ratio of hyperedges with low weight in one iteration.

occurrence frequency in documents after stemming. Also, an image is represented with a vector of histograms of 4,022 visual words extracted by SURF [14]. Then, values of each modality are converted to three-level values from 0 to 2 since hypernetwork models can deal with discretized data. Data are divided into a training set with 884 documents and 7,555 images and a test set consisting of 99 documents and 845 images for article retrieval. Table 1 shows the parameter setting to train layered hypernetworks.

5.2 Experimental Results

We evaluate the similarity of cross-modal associative generation by comparing generated text terms and visual words with text and image keywords in the given query. To evaluate the similarity, we define two measures in this paper. The first measure is ratio of correctness (RC). Referring a set whose elements are text terms and visual words which constitute a document and an image in an article to an original set, we generate text terms or visual words as same amount as the size of the original set. Then we compare a generated textual or visual set with the original set when partial text terms and visual words are given. RC is defined as follow:

$$RC = \frac{\text{\# of generated keywords same to keywords in an original set}}{\text{\# of generated text (image) keywords}} . \tag{15}$$

According to (15), RC can have a value from 0 to 1. The second measure is context score (CS) which are based on pair-wise co-occurrence of all text terms and visual words with non-negative value in documents and images of article data. To obtain CS, we define a measure of pair-wise co-occurrence for the *i*-th and *j*-th keyword as follow:

$$m_{ij} \begin{cases} \sum_{n=1}^{N} \frac{x_i^{(n)} \times x_j^{(n)}}{\left\{(x_i^{(n)})^2 - (x_j^{(n)})^2\right\}^2 + 1} & (i \neq j) \\ C & (i = j) \end{cases} , \tag{16}$$

where x_i and x_j is the value whose indices are *i* and *j* in the *n*-th data sample $\mathbf{x}^{(n)}$, *N* is the size of data set, and *C* is a arbitrary constant. Then, CS is defined as follow:

$$CS = \frac{1}{|G|} \sum_{i,j} m_{ij} , \tag{17}$$

where |*G*| is the size of set of generated text terms or visual words. The different point of CS from RC is that CS reflects the contexts of relationships between generated keywords. Although RCs of two generated sets are same, CSs may be different each other dependent on the co-occurrence frequency of wrongly generated keywords. Figure 4 and 5 are the result of text-to-image generating visual words and image-to-text generating text terms for all training data when a few text terms and visual words are given as a query. Figure 4 shows average RC and CS of generated text terms by image-to-text generation for 889 documents. Cross-modal queries can improve more 40% point of accuracy of the generation of text terms related to given queries

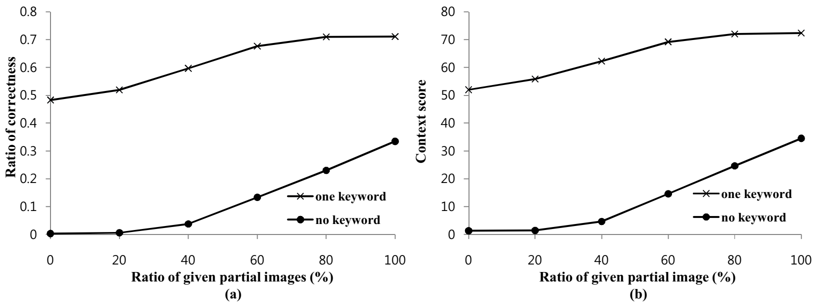

Fig. 4. Average RC (a) and CS (b) of generated visual words by image-to-text generation

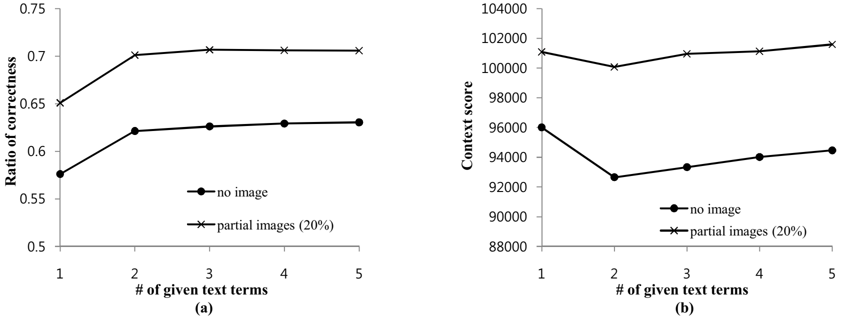

Fig. 5. Average RC (a) and CS (b) of generated keywords by text-to-image generation. Scale of context score of text to image generation is much larger than one of text-to-image generation since the size of image data is approximately ten times and non-zero variables in histogram vector of images are much more than in TF-IDF vector of documents.

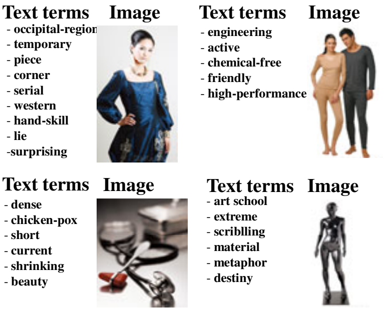

Fig. 6. Articles whose text terms are generated perfectly with given one text term and 20% of visual words in the article

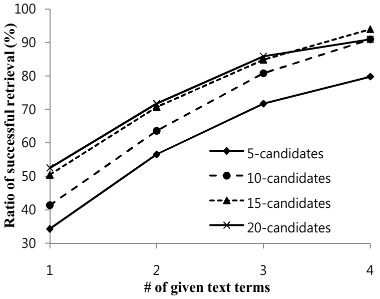

Fig. 7. Ratio of successful retrievals for test data set as the number of given text terms increases

compared with text query only. According to Figure 4, when the same amount of text terms is given, the similarity score of generated text terms get higher as information of given image increase. Also, without any text keyword query, text terms in the original set can generated with partial images only. Figure 5 presents average RC and CS of generated visual words by text-to-image generation for 884 images among training images. Same as Figure 4, multimodal information increases two scores compared with image input only. Dissimilar to image-to-text generation, RCs are saturated when more than two text terms are given. In addition, CSs show different patterns from image-to-text generations. It is the reason that an article consists of one document and several images so that image information is more important than text information. Figure 6 shows four pairs of the set of text terms and an image of articles whose RCs are 1 when one text terms and 20% of visual words in the article are given as a query. We can generate text terms and retrieve the article with small part of information by cross-modal associative generation. Figure 7 presents the ratio of successful article retrieval when partial text terms of a data are given for test data set using trained layered hypernetwork. In this study, article retrieval is considered to be successful when candidates include the test article whose text terms and visual words are given as a query. According to Figure 7, with both more than two text terms and half of image, the article which a user wants can be included over 90% when the size of candidates is 20.

6 Concluding Remarks

In this paper, we propose LHNs for cross-modal associative learning and a method to generate visual and textual keywords based on text-to-image and image-to-text cross-modal inference with LHNs for given multi-modal queries. Experimental results show that it is possible to generate keywords based on cross-modal association of inter-modalities. Also, multimodal queries improve the similarity of generated keywords compared with uni-modal ones. In addition, we show that proposed model and method can be applied to an articles retrieval system. As future works, we will apply the cross-modal associative keyword generation method to various problems such as auto-annotation for unlabeled images as well as multimodal information retrieval.

Acknowledgements

This work was supported in part by IT R&D Program of MKE/KEIT (KI002138, MARS), in part by NRF Grant of MEST (314-2008-1-D00377, Xtran), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2010-0017734), and in part by the BK21-IT program funded by Korean Government (MEST).

References

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

[2] Goh, K.-S., Chang, E.Y., Lai, W.-C.: Multimodal concept-dependent active learning for image retrieval. In: Proc. of the 12th Annual ACM International Conference on Multimedia (MM 2004), pp. 564–571 (2004)

[3] Simon, I., Snavely, N., Seitz, S.M.: Scene Summarization for Online Image Collections. In: Proc. of 11th IEEE International Conference on Computer Vision, ICCV 2007 (2007)

[4] Ha, J.-W., Kim, B.-H., Kim, H.-W., Yoon, W.C., Eom, J.-H., Zhang, B.-T.: Text-to-image cross-modal retrieval of magazine articles based on higher-order pattern recall by hypernetworks. In: Proc. of the 10th International Symposium on Advanced Intelligent Systems (ISIS 2009), pp. 274–277 (2009)

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

[6] Fuster, J.M., Bodner, M., Kroger, J.K.: Cross-modal and cross-temporal association in neurons of frontal cortex. Nature 405, 347–351 (2000)

[7] Snoek, C.G.M., Worring, M.: Concept-based video retrieval. Foundations and Trends in Information Retrieval 2(4), 215–322 (2009)

[8] Yan, R., Hauptmann, A.G.: A review of text and image retrieval approaches for broadcast news video. Information Retrieval 10(4-5), 445–484 (2007)

[9] Li, D., Dimitrova, N., Li, M., Sethi, K.: Multimedia content processing through cross-modal association. In: Proc. of the 11th Annual ACM International Conference on Multimedia (MM 2003), pp. 604–611 (2003)

[10] Ferecatu, M., Boujemaa, N., Crucianu, M.: Semantic interactive image retrieval combining visual and conceptual content description. Multimedia Systems 13, 309–322 (2008)

[11] Yakhnenko, O., Honavar, V.: Annotating images and image objects using a hierarchical dirichlet process model. In: Proc. of the 9th International Workshop on Multimedia Data Mining in ACM SIGKDD 2009, pp. 1–7 (2009)

[12] Quek, F., McNeil, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Trans. on Computer-Human Interaction 9(3), 171–193 (2002)

[13] Christoudias, C.M., Saenko, K., Morency, L.-P., Darrell, T.: Co-Adaptation of audio-visual speech and gesture classifiers. In: Proc. of the 8th International Conference on Multimodal Interfaces, pp. 84–91 (2006)

[14] Bay, H., Tuytelaars, T., Gool, T.V.: Surf: Speed up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)