하이퍼네트워크 모델을 이용한 비전-언어 크로스모달 연관정보 추출

Extraction Analysis for Crossmodal Association Information using Hypernetwork Models

허민오, Min-Oh Heo*, 하정우, Jung-Woo Ha**, 장병탁, Byoung-Tak Zhang***

요약 하나의 컨텐츠를 위해 동영상, 이미지, 소리, 문장과 같은 하나 이상의 모달리티로 전달하는 멀티모달 데이터가 증가하고 있다. 이러한 형태의 자료들은 잘 정의되지 않은 형태를 주로 가지기 때문에, 모달리티 간의 정보가 명백히 표현되지 못하는 경우가 많았다. 그래서, 본 연구에서 저자들은 자연계를 다루는 다큐멘터리 동영상 데이터를 이용하여 비전-언어 간의 상호 연관정보인 크로스모달 연관정보를 추출하고 분석하는 방법을 제시하였다. 이를 위해 정글, 바다, 우주의 세 가지 주제로 구성된 다큐멘터리로부터 이미지와 자막의 조합으로 이루어진 데이터를 모은 후, 그로부터 시각언어집합과 문장언어집합을 추출하였다. 분석을 통하여, 이 언어집합들간의 상호 크로스 모달 연관정보를 통해 생성된 다른 모달리티 데이터가 의미적으로 서로 관련이 있음을 확인할 수 있었다.

Abstract Multimodal data to have several modalities such as videos, images, sounds and texts for one contents is increasing. Since this type of data has ill-defined format, it is not easy to represent the crossmodal information for them explicitly. So, we proposed new method to extract and analyze vision-language crossmodal association information using the documentaries video data about the nature. We collected pairs of images and captions from 3 genres of documentaries such as jungle, ocean and universe, and extracted a set of visual words and that of text words from them. We found out that two modal data have semantic association on crossmodal association information from this analysis.

핵심어: *Higher-order pattern analysis, Multimodal Data, Crossmodal Association, Vision-language analysis*

본 논문은 2008 년 교육과학기술부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구(KRF-2008-314-D00377)이며, 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅(MEC) 과제와 교육과학기술부의 BK21-IT 지원 및 과학기술부의 국가지정연구실사업(NRL)으로 일부 지원되었음. (No. M10400000349-06J0000-34910)

*주저자 : 서울대학교 전기컴퓨터공학부 박사과정 e-mail: moheo@bi.snu.ac.kr

**공동저자 : 서울대학교 전기컴퓨터공학부 박사과정 e-mail: jwha@bi.snu.ac.kr

***교신저자 : 서울대학교 전기컴퓨터공학부 교수; e-mail: btzhang@bi.snu.ac.kr

1. 서론

멀티모달 데이터 (Multimodal Data)는 텍스트, 이미지, 소리, 동영상과 같은 하나 이상의 모달리티로 정보를 전달하는 형식을 의미한다. 멀티모달 데이터는 표현하고자 하는 바를 위하여 여러 모달리티의 정보를 지니고 있으므로, 언어 처리, 이미지 분석과 같은 기존의 단일 모달리티 처리기술을 기반으로 하여 디지털 멀티미디어 데이터의 핵심 요소인 텍스트, 이미지, 소리, 동영상 사이의 연관관계를 다루는 기반이 될 수 있다. 이러한 특성을 이용하여 보다 나은 음성 인식, 동영상 내의 행동분석, 주의집중 모델링, 멀티미디어 검색, 이미지 검색, 동영상 스크립팅과 같은 다양한 분야에 응용하려는 연구들이 진행되고 있다 [1-3] .

크로스 모달 연관성 (Crossmodal Association)은 여러 모달리티 간의 상호 연관성을 의미한다. 인간이 지니고 있는 시각, 청각, 촉각을 비롯한 여러 모달리티 간의 연관성에 대한 연구결과 [4, 5] 뿐만 아니라, 멀티모달 데이터 상의 여러 모달리티 간의 연관성을 이용하면, 기존의 정보 검색 기술의 성능을 보다 향상시킬 수 있다 [1, 2]. 최근 늘어나고 있는 멀티미디어 데이터에 대한 정보검색의 필요성이 크게 증가함에 따라, 이미지와 텍스트 사이의 연관관계를 분석하여 이미지에 키워드를 태깅하는 기법인 자동 태깅에 관한 연구 [1-3, 6-11]가 진행되어 오고 있으며, 특히, 모달리티 상호 검색에 중점을 둔 접근은 [10-11]을 들 수 있다. 이들은 데이터 인스턴스 사이의 유사도 함수에 기반한 접근을 하고 있기 때문에, 크로스 모달 연관정보를 직접 표현하지 못하므로 한계가 존재한다.

크로스 모달 연관정보를 직접 표현하기 위해서는 텍스트의 의미단위가 단어인 것처럼, 이미지의 의미단위를 반영한 프리미티브 (Primitive)로의 표현이 필수적이다. 이를 위하여, 이미지 데이터 집합 내에서 의미단위를 추출하여 새로운 추상적 표현단계를 생성하는 데에 프리미티브가 되는 시각언어 (Visual word)를 추출하는 방법을 이용한다.

모달리티를 함께 학습하고 하나의 모달리티 쿼리가 주어졌을 때 다른 모달리티의 정보를 인출할 수 있는 모델이 필요하다. 이러한 모델로 가중치를 부여한 하이퍼그래프 (Hypergraph)를 다루는 확률 그래프 모델인 하이퍼네트워크 모델 (Hypernetwork Model) [12]을 사용하기로 한다.

이 모델은 각각의 데이터의 특성 벡터들로부터 선택된 특성들의 조합을 연결하여 그래프 구조로 표현한다. 하나의 그래프는 특성을 정의된 확률에 따라 랜덤하게 선택하여 그래프의 구조를 생성하므로 랜덤 그래프 (Random Graph)가 되며, 여기에 가중치를 부여함으로써 특성-특성간의 연관관계와 특성-데이터간의 연관관계를 긴밀하게 분석하고 관찰하는 것이 가능하다[12]. 따라서 이 모델을 이용하여 이미지와 텍스트 사이의 연관정보를 생성하고 분석할 것이다.

이후의 구성은 다음과 같다. 2 장에서는 이 문제들을 다루기 위한 모델로서 하이퍼네트워크 모델을 소개하고 특성들을 간단히 기술할 것이다. 3 장에서는 어떻게 주어진 단어로부터 이미지를 검색할 것이며, 이미지가 주어졌을 때, 어떻게 키워드를 생성할 것인지에 대한 방법을 제시할 것이다. 4 장에서는 데이터 획득 및 생성과 실험 결과에 대해 설명하고, 5 장에서 결론을 맺을 것이다.

2. 관련 연구

2.1 하이퍼네트워크 모델

하이퍼그래프 모델 (hypergraph model)은 모델 내의 에지 (edge)가 3 개 이상의 버텍스 (vertex)를 동시에 연결할 수 있는 하이퍼에지로 구성된 그래프 모델이다. 그러므로 하이퍼그래프는 인자들간의 복잡한 고차원적 인과관계로 표현되는 실세계의 문제를 모델링 하는데 있어서 기존의 그래프 모델에 비해서 표현성의 측면에서 장점을 가지며 이러한 특성으로 인해 기존의 그래프 모델로서 표현되던 문제를 하이퍼그래프 모델 문제로 확장하는 것은 자연스럽게 진행할 수 있다.

일반적으로 하이퍼그래프 모델을 수식으로 표현하면 V 를 버텍스의 집합, E 를 하이퍼에지의 집합, W 를 하이퍼에지의 가중치 (weight)의 집합이라 할 때 하이퍼그래프모델 G는 $G = (V, E, W)$로 표현된다.

하이퍼네트워크는 하이퍼그래프의 한 종류로서, 하이퍼네트워크 모델에서 버텍스는 데이터를 구성하는 인자 (attribute)를 – 구체적으로는 인자의 인덱스와 인자 값의 쌍 – 의미하고 하이퍼에지가 인자들간의 임의의 조합을 표현한다. 그림 1 은 하이퍼네트워크 모델을 표현하고 있으며 그림 내에서 다양하게 표현된 다각형이 하이퍼에지를 의미한다.

하이퍼네트워크를 수식으로 표현하면 하이퍼그래프와 마찬가지로 하이퍼네트워크 H 는 $H=(V, E, W)$로 표현된다

그리고 하이퍼에지를 구성하는 인자의 수를 하이퍼에지의 차수 (cardinality) 혹은 오더 (order)라 정의

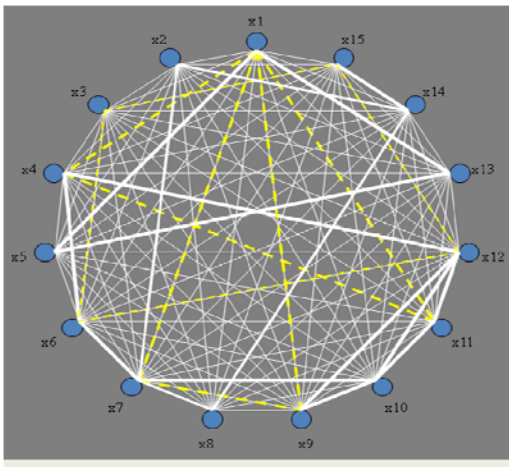

그림 1. 하이퍼네트워크 모델

하며 k 오더 하이퍼에지로만 구성된 하이퍼네트워크를 k-uniform 하이퍼네트워크라고 정의한다. 하이퍼에지가 인자들간의 조합 형태를 띠게 되므로 이는 인자들로 구성된 conjunctive normal forms (CNFs) 으로 이해될 수 있다. 또한 같은 원리로 하이퍼네트워크는 하이퍼에지들의 disjunctive normal forms (DNFs)로 인식될 수 있다[13]. 그러므로 하이퍼네트워크는 수많은 CNF 로 구성된 연관규칙 (association rule) 집합의 성격을 갖게 된다.

일반적인 하이퍼네트워크의 학습 및 모델링 과정은 그림 2 와 같다. 먼저 주어진 데이터는 훈련 (training), 검증 (validation), 테스트 (test) 집합으로 분할된다. 훈련데이터는 하이퍼에지를 생성하고 모델을 학습하는 데 사용된다. 그리고 검증 데이터는 매 반복학습 (iterations) 마다 학습된 모델의 성능을 검증하는 데 사용되며 검증 데이터에 대해 최소의 에러를 갖는 모델을 선택하여 테스트 데이터에 대하여 일반화 성능을 측정한다.

하이퍼에지는 훈련데이터로부터 무작위 샘플링을 통해 생성되며 생성된 하이퍼에지들이 하이퍼네트워크를 구성한다. 구성된 하이퍼네트워크에 대해서 하이퍼에지의 가중치는 훈련데이터 샘플이 주어질 때 하이퍼에지를 구성하는 버텍스들과 데이터 샘플 내에 있는 버텍스들에 해당하는 인자값이 같은가를 확인한다. 이 때 모든 하이퍼에지 내의 모든 버텍스들의 값과 데이터 샘플 내의 인자값이 동일한 경우 하이퍼에지의 클래스 값과 데이터 샘플의 클래스 값이 같은 경우 가중치가 증가하게 되며, 클래스 값이 동일하지 않은 경우 감소하게 된다. 모든 훈련데이터가 하이퍼네트워크 내의 모든 하이퍼에지에 대해 가중치 산출 및 갱신이 완료되면 검증데이터를 이용해서 분류성능을 측정한다.

그리고 일정 횟수 동안 학습을 반복한 후 테스트 데이터를 이용해 일반화 분류 성능을 측정한다. 하이퍼네트워크의 자세한 학습 및 모델링 과정은 Zhang 의 연구 [12] 에 설명되어 있다.

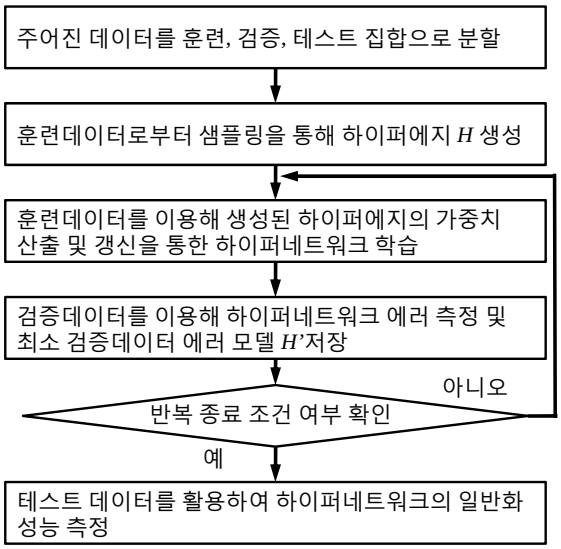

그림 2. 하이퍼네트워크 모델의 학습 과정

단계	사용한 알고리즘
국소 특징점 추출	MSER detector
기술	SIFT descriptor
시각언어 요약	k-means clustering

표 1. 시각 언어 추출에 쓰인 방법

그림 3. 수집된 이미지 중 국소추출자의 적용 예

2.2 시각언어 (Visual Words)

단어 단위의 텍스트와 함께 이미지를 다루기 위해서는 이미지의 의미단위 표현을 할 수 있는 프리미티브가 필요하다. 이를 위하여, 최근 이미지 연구에서 빈번하게 쓰이고 있는 이미지 표현 방법인 시각언어 (Visual Words)를 추출하여 사용하였다. (표 1 참고)

시각언어를 구하는 일반적인 방법은 크게 3 단계로 나누어져 있다. 첫 번째 단계는 이미지 상에서 특정 기준에 따라 관심대상이 되는 지역별 특징점들을 추출해내는 추출자 (local feature detector)를 사용하는 것이다. 본 연구에서 쓰인 국소 특징점 추출자는 image intensity 를 기준으로 극소값과 극대값의 주위 영역을 추출하는 MSER (Maximally Stable Extremal Regions) [14]이다.

두 번째 단계는 관심대상인 추출된 이미지 조각들을 잘 정의한 인자 공간 (feature space)에 할당하는 것이다. 추출된 국소 특징점 근처의 이미지를 하나의 벡터로 변환하는 기술자 (descriptor)를 이용하여 양수 차원의 실수공간인 인자 공간에 매핑 한다. 본 연구에서는 주어진 이미지 조각을 기울기 (Gradient)의 히스토그램을 이용하여 128 차원으로 기술하는 SIFT (Scale Invariant Feature Transform) 기술자 [15]를 이용한다.

세 번째 단계는 인자 공간 상에 매핑된 이미지 조각들을 미리 정해둔 시각언어의 단어 개수로 요약하는 것이다. 이미 많이 알려져 있는 많은 클러스터링 알고리즘 중에 하나를 활용하여 클러스터링을 수행하고, 미리 정해둔 시각언어의 단어 수만큼의 클러스터 중심을 찾는 방법을 이용하여 시각언어를 구한다. 본 연구에서는 클러스터링 알고리즘으로 k-means clustering 방법을 이용하고, k 값은 1000 을 썼다.

3. 크로스모달 연관 정보 추출방법

3.1 크로스모달리티 생성

멀티모달 데이터는 여러가지 모달리티 정보가 인자로 표현될 수 있으며 동영상 데이터의 경우 이미지와 텍스트

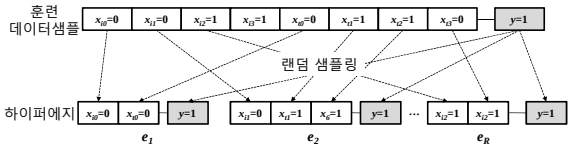

그림 4. 이미지와 텍스트 복합정보로 표현된 데이터 샘플로부터 하이퍼에지를 생성하는 과정

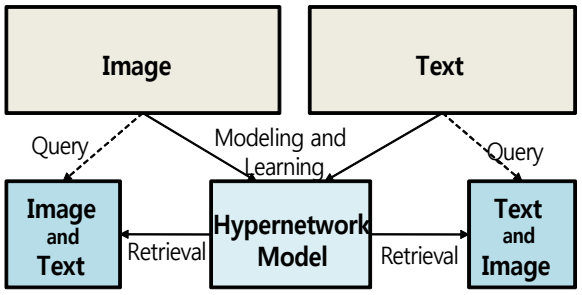

그림 5. 비전-언어 모달리티 간 모델링 및 크로스모달 연관정보 생성과정

그리고 음성의 세 가지 모달리티로 표현될 수 있다. 본 연구에서는 크로스 모달 연관정보를 다루기 위하여 하이퍼네트워크 모델을 이용하여 음성을 제외한 이미지와 텍스트 정보간의 크로스 모달 학습을 구현하고, 학습된 모델로부터 이미지에서 텍스트, 텍스트에서 이미지 사이의 연관성을 분석하는 문제를 도입한다.

이미지와 텍스트 모달리티 정보를 인자로 표현하기 위해 각 인자값 x_i, x_t 는 bag-of-words 모델과 같이 데이터 내에서 시각언어와 해당 키워드의 존재여부를 {0, 1}을 이용하여 이진 (binary) 형태로 표현되었다.

그림 4 는 이미지 정보와 텍스트 정보로 구성된 데이터 샘플로부터 하이퍼에지를 생성하는 과정을 설명하고 있다. 또한 각각의 서로 모달리티로부터 정보를 생성하는 과정은 그림 5 와 같이 도식화 될 수 있다.

주어진 데이터에서 이미지 정보를 I, 텍스트 정보를 T, 하이퍼에지의 가중치를 W 로 두면, I 와 T 가 각각 p 개와 q 개의 인자로 구성되어 있을 때 I와 T는 각각

$$I = \{x_{i1},\ x_{i2},\ \cdots,\ x_{ip}\} \tag{1}$$

$$T = \{x_{t1},\ x_{t2},\ \cdots,\ x_{tq}\} \tag{2}$$

으로 표현가능하며 이미지와 텍스트 정보로 표현된 하이퍼네트워크의 joint probability 는

$$P(I,\ T|\ W) \ = \ P(x_i,\ x_t|\ W) \ = \ P(x|\ W) \tag{3}$$

가 된다. 여기에서 이미지 정보가 주어질 경우 연관성이 높은 텍스트 키워드를 찾는 과정을 수식화하면

$$P(T \mid I) = \frac{P(I,T)}{P(I)} = \frac{P(x_{i1}, x_{i2}, ...x_{ip}, x_{t1}, x_{t2}, ..., x_{tq})}{P(x_{i1}, x_{i2}, ..., x_{ip})} \tag{4}$$

으로 표현된다. 또한 텍스트 정보로부터 연관성이 높은 이미지 또는 이미지 조각을 찾는 과정을 수식화하면

$$P(I \mid T) = \frac{P(I,T)}{P(T)} = \frac{P(x_{i1}, x_{i2}, ...x_{ip}, x_{t1}, x_{t2}, ..., x_{tq})}{P(x_{t1}, x_{t2}, ..., x_{tq})} \tag{5}$$

으로 표현된다.

3.2 크로스모달 학습

하이퍼네트워크에서의 크로스모달 학습은 위의 그림 4 에서와 같이 서로 다른 여러 개의 모달리티 정보를 하나의 하이퍼에지에 샘플링하고 생성된 하이퍼에지를 학습함으로써 실현된다. 본 연구에서는 크로스모달 학습을 통해 이종 모달리티 간 데이터 재생성을 목적으로 하고 있기 때문에 일반적인 하이퍼네트워크의 학습모델[12]과는 다른 방법이 적용되었으며 이 연구에서 적용된 학습 방법은 다음 그림 6 과 같다. 학습과정에서 인자 값이 1 인 것만 샘플링 하여 에지를 구성하는 이유는 데이터의 특성상 인자 차원이 매우 큰데 비해 상대적으로 데이터 샘플의 수가 작고 또한 각 시각언어 및 키워드 별 1 값이 상대적으로 0 에 비해 빈도가 매우 작기 때문이다. 또한 의미상으로 다른 두 개의 시각언어 및 키워드가 서로 동시에 존재하는 경우를 강조해서 학습시키기 위함이다.

(1) 값이 1 인 인자에 대하여 각 모달리티 별로 최소 하나 이상의 인자를 샘플링하여 하이퍼에지 생성. 또한 하이퍼에지 별로 시각언어 및 텍스트 키워드의 카운트 변수 포함시킴
$e_i = \{x^{(k)}, c_{vw}, c_{tw}, w_i\}$: k-오더 하이퍼에지

(2) 하이퍼에지 내의 각 인자 값과 훈련 데이터 내에서의 인자 값의 동일여부를 비교, 즉 훈련 데이터 내에 하이퍼에지가 포함한 시각언어 혹은 키워드가 존재하는지 비교

(3) 하이퍼에지 내의 시각언어 혹은 텍스트 키워드가 훈련 데이터 샘플에 존재하는 경우 에지 내의 각각의 카운트를 증가시킴

(4) 모든 훈련 데이터와 하이퍼에지의 값 비교 후 하이퍼에지의 가중치 값 산출
$w_i = c_{vw} + c_{tw}$

(5) 측정된 가중치 값을 기준으로 하이퍼에지의 하위 일정 비율을 제거하고 제거된 하이퍼에지 수에 비례하게 (1) 단계를 정해진 수만큼 반복 실행

그림 6. 이미지 및 키워드 생성을 위한 하이퍼네트워크 학습 방법

4. 실험결과

정글, 바다, 우주를 다루는 세 장르의 동영상 다큐멘터리에서 자막이 나타나는 시점을 기준으로 하여 화면과 자막을 저장하고 짝을 지었다. 이와 같은 작업으로 그림 7 과 같은 1118 쌍의 자막과 이미지를 얻었다. 화면 이미지 집합에서 시각언어를 추출하고, 텍스트 단어와 함께 하이퍼네트워크 모델을 이용하여 학습을 수행하였다.

시각언어와 텍스트 단어는 서로 장르 별로 나눠지지 않고 겹쳐지는 것들이 존재한다. 학습을 통해 얻은 시각언어와 텍스트 단어는 표 2 와 그림 8 과 같이 장르별로 다소 다른 분포를 가진다. 데이터 인스턴스 당 시각언어의 수가 데이터 인스턴스 당 텍스트 단어에 비하면 194.36 과 9.04 로 20 배 이상 많다. 이것은 하나의 시각언어는 개념의 일부를 나타내는 데에 참여할 수는 있어도 개념 하나의 의미를 담기에는 다소 부족하다는 점에 기인한다.

장르	데이터 개수	시각언어 수	텍스트 단어 수
정글	440	968	1336
바다	154	969	622
우주	524	969	1237
총 개수	1118	1000	2440
인스턴스 당 평균	–	194.36	9.04

표 2. 수집된 데이터 요약

그림 7. 우주 장르 중의 두 데이터 쌍의 예

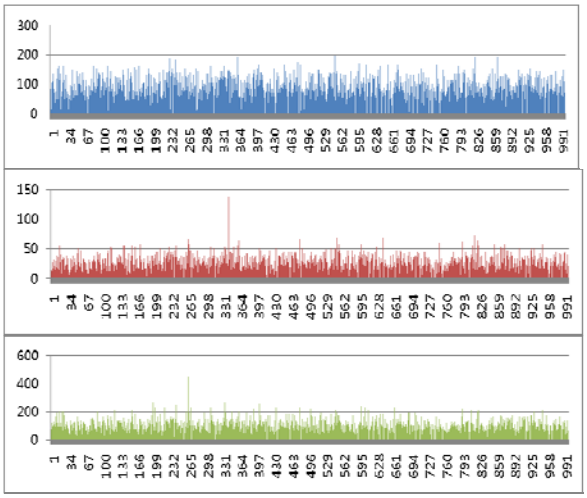

그림 8. 시각언어 인덱스에 따른 빈도 (위에서부터 정글, 바다, 우주)

4.1 입력이 없을 시 연관정보 추출

입력 값이 없을 경우의 연관정보 추출 결과는 학습된 하이퍼네트워크 모델 내의 하이퍼에지 중에서, 가중치가 높은 하이퍼에지를 이용하여 얻을 수 있다. 각 모달리티 별 오더는 2 로 샘플링하여 4-오더 하이퍼에지를 생성하고 3.2 절의 과정에 따라 학습하였다. 그림 9 는 쿼리가 없을 경우, 하이퍼에지의 웨이트가 가장 큰 결과이다.

일반적으로 언어에서의 단어 빈도는 of, the, a, is, and, to, in 와 같은 관사, be 동사, 접속사처럼 의미를 부여하기보다는 기능을 수행하는 단어들이 빈번하게 쓰이기 때문에, 아무런 입력 값이 없을 경우 그림 10 과 같이 이러한 텍스트 단어들이 포함된 joint probability 가 높게 나타난다. 시각 언어에도 유사한 성질을 지니는 것들이 존재한다 [16]. 그림 9 의 결과는 이미지 내의 객체보다는 배경을 나타내는 데에 참여하는 소수의 시각언어가 빈번하게 존재함을 보이고 있다.

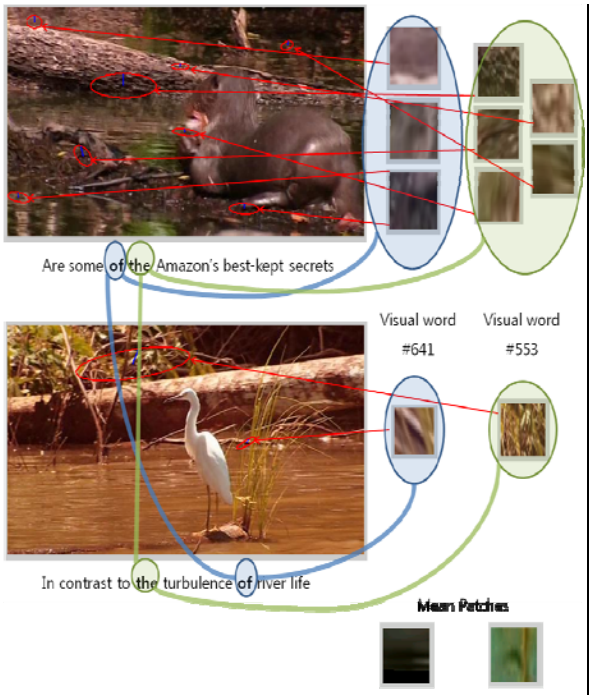

그림 9. 입력이 없을 시, 스코어가 가장 높은 하이퍼에지가 나타내는 연관정보. 시각단어 641, 553 과 매칭되는 텍스트단어 of, the 가 나타나는 두 개의 이미지와 문장 쌍의 예이다. 왼쪽은 원본이미지와 문장을 표시하고 오른쪽은 시각 언어와 원본 이미지 내에 나타난 시각언어의 형상을 표시한 것이다. 연결선들이 어떻게 시각언어와 텍스트 단어가 연결되는지 보여주고 있다.
Mean Patches 는 두 시각단어 641,553 이 데이터 집합 전체에서 나타나는 이미지 조각들의 평균 이미지를 표시한 것이다.

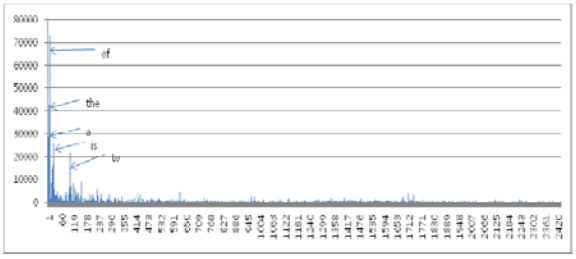

그림 10. 하이퍼네트워크 상의 텍스트 단어 인덱스별 웨이트

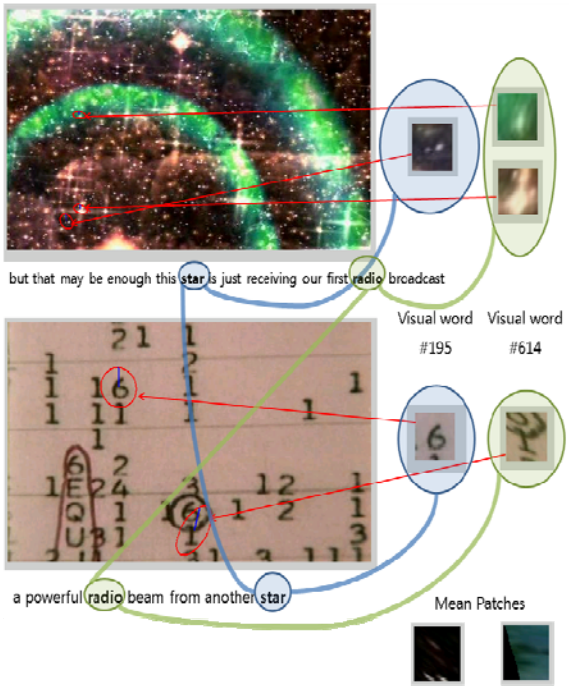

그림 11. 입력이 주어졌을 때, 스코어가 가장 높은 하이퍼에지가 나타내는 연관정보. 시각단어 195, 614 와 매칭되는 텍스트단어 star, radio 가 나타나는 두 개의 이미지와 문장 쌍의 예이다. 왼쪽은 원본이미지와 문장을 표시하고 오른쪽은 시각 언어와 원본 이미지 내에 나타난 시각언어의 형상을 표시한 것이다. 연결선들이 어떻게 시각언어와 텍스트 단어가 연결되는지 보여주고 있다.
Mean Patches 는 두 시각단어 195,614 가 데이터 집합 전체에서 나타나는 이미지 조각들의 평균 이미지를 표시한 것이다.

4.2 입력값에 따른 연관정보 추출

입력 값이 주어진 경우의 연관정보 추출 결과는 학습된 하이퍼네트워크 모델 내의 하이퍼에지 중에서, 3.1 절에서 설명한 조건부 확률에 따른 가중치가 높은 하이퍼에지를 이용하여 얻을 수 있다. 각 모달리티 별 오더는 2 로 샘플링하여 4-오더 하이퍼에지를 생성하고 3.2 절의 과정에 따라 학습하였다. 입력 값은 텍스트 데이터 집합 내에서 두 개의 명사를 택하였다. 그림 11 은 입력 값이 'star', 'radio' 일 경우의 하이퍼에지 웨이트가 가장 큰 결과이다.

시각언어는 이미지 조각 내의 선의 방향과 밝기를 비롯한 여러 정보들이 종합되어 있는 것이기 때문에, 텍스트 단어의 의미와 인간이 인식하는 시각언어의 내용이 일관되지 않는 경우도 나타난다. 하지만, 다수의 텍스트 언어와 다수의 시각언어가 함께 다루어질 경우에는 보다 의미를 많이 포함하고 있을 것이다.

5. 결론

크로스모달 연관정보는 멀티모달 데이터의 양이 증가하면서 분석의 중요성이 높아지고 있다. 본 논문에서는 하이퍼네트워크 모델을 통하여 비전-언어 간의 크로스모달 연관정보를 추출하는 방법을 소개하고 실험을 수행하였다. 입력이 없을 시에 연관정보를 추출해보았을 때, 가장 빈번한 패턴을 보이는 시각언어, 텍스트 언어의 조합을 찾을 수 있었으며, 입력값에 따라 의미적 연관을 발견할 수 있었다. 상호 크로스 모달 연관정보를 통해 생성된 다른 모달리티 데이터가 의미적으로 서로 관련이 있음을 확인할 수 있었다.

참고문헌

[1] P. Maragos, A. Potamianos, and P. Gos, "Multimodal Processing and Interaction: Audio, Video, Text (Multimedia Systems and Applications)", Springer, Berlin, Germany, 2008.

[2] D. Li, N. Dimitrova, M. Li, and K. Sethi, "Multimedia Content Processing through Cross-Modal Association", Proceedings of the 11th ACM International Conference on Multimedia, pp. 604~611, 2003.

[3] F. Murtagh, A. Ganz, and S. Mckie, "The structure of narrative: The case of film scripts", Pattern Recognition 42, pp. 302~312, 2009

[4] J. M. Fuster, M. Bodner, and J. K. Kroger, "Cross-modal and Cross-temporal Association in Neurons of Frontal Cortex", Nature, Vol. 405, pp. 347~351, 2000

[5] Y.-D. Zhou, and J. M. Fuster, "Visuo-Tactile Cross-modal Associations in Cortical Somatosensory Cells", Proceedings of the National Academy of Sciences, Vol. 97, No. 17, pp. 9777~9782, 2000

[6] E. Chang, G. Kingshy, G. Sychay, and G. Wu, "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines", IEEE Transactions on Circuits and Systems for Video Technology, Vol.13, No.1, IEEE Circuits and Systems Society, pp. 26~38, 2003

[7] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures", Journal of Machine Learning Research, Vol. 3, pp. 1107~1135, 2003

[8] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, pp. 119~126, 2003

[9] M. Zhu, and A. Badii, "Semantic-associative Visual Content Labelling and Retrieval: A Multimodal Approach", Image Communication, Vol. 22, Issue 6, pp. 569~582, 2007

[10] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic Multimedia Cross-modal Correlation Discovery", Proceedings of the 10th ACM SIGKDD Conference on Knowledge discovery and data mining, Association for Computing Machinery, pp. 653~658, 2004

[11] J. Liu, B. Wang, H. Lu, and S. Ma, "A Graph-based Image Annotation Framework", Pattern Recognition Letters, Vol. 29, Elsevier, pp. 407~415, 2008

[12] B.-T. Zhang, "Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory", IEEE Computational Intelligence Magazine, Vol. 3, Issue 3, pp. 49~63, 2008

[13] B.-T. Zhang, and H.-Y. Jang, "A Bayesian Algorithm for in vitro Molecular Evolution of Pattern Classifiers", Lecture Notes in Computer Science 3384, pp. 458-467, 2005

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions", Proceedings of the 13rd British Machine Vision Conference, pp. 384~393, 2002

[15] D. G. Lowe, "Object Recognition from Local Scale-invariant Features", Proceedings of the 7th IEEE International Conference on Computer Vision, Vol. 2, pp. 1150~1157, 1999

[16] J. Sivic, and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos", Proceedings of the 9th IEEE International Conference on Computer Vision, Vol. 2, pp. 1470~1477, 2003.