멀티 모달 개념 망과 언어 모델을 이용한 이미지 설명문 생성

김경민 ○, 하정우, 이범진, 장병탁

서울대학교 컴퓨터공학부

{kmkim, jwha, bjlee, btzhang}@bi.snu.ac.kr

Generating image descriptive sentences via multi-modal concept networks and language models

Kyung-Min Kim, Jung-Woo Ha, Bum-Jin Lee, Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

시맨틱 네트워크에서와 같이 개념은 인공지능 분야에서 지식을 표현하는 수단으로 사용되어왔고 이러한 개념들은 주로 대형 코퍼스나 위키피디아와 같은 언어 형태의 데이터로부터 학습 및 표현되었다. 그러나 최근 비디오와 같은 멀티 모달 데이터가 급격히 늘어남에 따라 데이터에 포함된 지식의 양도 급격히 증가하게 되었으며 이에 따라 이미지와 같은 언어 형태 외의 멀티 모달 데이터로부터 개념을 학습하고 표현할 수 있는 연구의 필요성이 대두되었다. 본 논문에서는 멀티 모달 하이퍼네트워크를 이용하여 전체 비디오로부터 단어와 이미지 패치로 표현되는 개념들 간의 고차 연관관계를 학습하고 학습된 결과로부터 멀티 모달 개념 망을 생성한다. 그리고 생성된 개념 망을 통해 비디오의 특정 화면 이미지를 설명할 수 있는 문장을 생성한다. 입력된 이미지 질의들은 멀티 모달 개념망을 이용한 모달리티 교차 질의 확장을 통해 단어 질의로 변환되고, 변환된 단어 질의들은 비디오의 모든 문장을 코퍼스로 학습한 SRILM 언어모델을 통해 이미지 설명 문장으로 생성된다. 실험결과로서 본 논문에서는 517분의 상영시간을 갖는 유아용 만화 비디오 'Maisy'로부터 멀티 모달 개념 망을 생성하고 생성된 개념 망과 특정 화면 이미지를 설명하는 다양한 문장을 생성하였다.

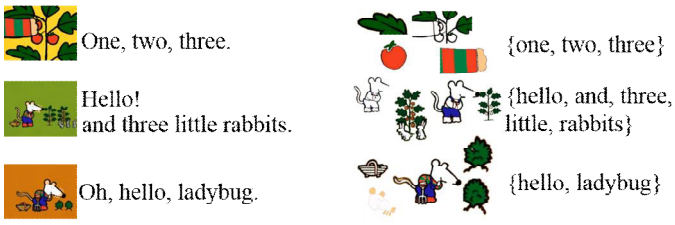
1. 서론

개념은 WordNet처럼 지식표현에 사용될 수 있다[1,2]. 이러한 지식 들은 실세계에 존재하는 비디오, 이미지, 텍스트 등을 통해 얻어지게 된다. 지금까지 인공지능에서 사용된 개념은 대부분 주어진 데이터가 변하지 않는 상황에서 단일 모달 데이터만 표현할 수 있었다. 최근, 비디오 데이터의 크기와 사용빈도가 증가함에 따라 멀티 모달 데이터로부터 지식을 획득하고 표현하는 방법이 중요성이 확대되고 있다. 이러한 멀티모달 지식 획득 및 표현 방법은 비디오 검색 및 요약에 활용 가능하며 비디오의 화면의 내용을 설명할 수 있는 문장을 생성하는 데 적용될 수 있다. 본 논문에서는 멀티 모달 하이퍼네트워크를 이용해 비디오 데이터를 학습하여 멀티 모달 개념망을 만들었다[3]. 하이퍼네트워크 모델은 확률기반 분산형 연관 메모리이며 점진적인 데이터 학습에 적합하다[4]. 또한 인지 모델에도 적용할 수 있다[5]. 그리고 비디오의 화면 이미지를 모달리티 교차 질의 확장을 통해 단어 질의로 변환한 뒤 이미지를 설명하는 문장을 만들어보도록 시도했다. 비디오 데이터로 유아용 만화 'Maisy'를 사용했다. 만화 비디오는 이야기 구성이 분명하고 이야기가 진행될수록 개념들의 관계가 계속 변한다는 점에서 평가에 유용했다. 그리고 언어 모델로는 SRI language model (SRILM)[6]을 사용했고 'Maisy' 만화 비디오의 에피소드 1부터 6까지 517분 분량의 모든 자막을 코퍼스로 학습했다.

2. 멀티 모달 개념 망 구축 알고리즘

2.1 데이터 전처리

비디오 데이터에서는 그림 1(a)와 같이 화면 이미지와 이에 해당하는 자막이 연속적으로 나타난다. 개념을 표현하기 위해서 각각의 화면 이미지는 그림 1(b)와 같이



(a) 캡쳐된 비디오 이미지 (b) 이미지 패치-단어 집합

그림 1.(a) 비디오 데이터는 이미지와 해당 자막의 연속으로 표현될 수 있다. (b) 이미지와 자막은 이미지 패치의 집합과 단어의 집합으로 변환된다.

MSER에 의해 추출된 이미지 패치의 집합으로 바뀌어졌고 패치들은 SIFT에 의해 히스토그램 벡터로 나타내어졌다[7,8]. 화면 이미지에 해당하는 자막은 단어의 집합으로 바뀌어졌고 이에 따라 비디오 데이터는 이미지 패치와 단어 집합들로 표현할 수 있게 되었다. 단어와 달리 이미지 패치들은 두 패치가 개념적으로 같은지 여부를 판단하기 위한 방법이 필요하다. 이를 위해 전체 이미지 패치들을 대상으로 히스토그램 벡터들 사이 L2-distance에 따라 *k*-means clustering을 실행했다.

2.2 멀티 모달 개념 망 구축 방법

하이퍼네트워크 모델은 메모리에 기반한 고차 확률 그래프 모델이며 하이퍼그래프 구조를 통해 모델을 표현한다[4]. 하이퍼그래프는 하나의 에지(하이퍼에지)가 두 개 이상의 노드들을 동시에 연결할 수 있는 그래프이다[3]. 하이퍼네트워크 모델에서 하나의 노드는 이미지 패치와 단어로 이뤄진 개념을 나타내고 하이퍼에지는 노드들 사이 관계를 나타낸다. 또한 하이퍼에지의 가중치는 관계의 강도를 반영한다[3]. 그리고 기존의 시맨틱 네트워크에서 에지는 개념들 사이의 'is-a', 'is a part of', 'lives in'과 같은 관계를 나타냈지만, 하이퍼네트워크 모델에서 에지는 통계적인 관계를 나타낸다. 이를 통해 개념들간의 근접도에 초점을 두게 된다[9].

하이퍼네트워크 모델을 구축하는 방법에 대해 간단히 설명하면 다음과 같다.

(1) 비디오의 모든 이미지 패치와 단어 집합에 대해서 랜덤 샘플링을 하고 일정한 수의 하이퍼에지를 생성한다.
(2) 비디오에서 이야기가 진행됨에 따라 새로 들어오는 이미지 패치와 단어 집합과 비교하여 일치 여부에 따라서 하이퍼에지의 가중치를 증가하거나 감소시킨다.
(3) 특정 역치 값 아래로 떨어지는 가중치를 갖는 하이퍼에지를 제거한다.
(4) 이야기가 계속 진행됨에 따라 이미지 패치-단어 집합이 하이퍼네트워크 노드 집합에 포함되고 하이퍼에지도 새로 생성된다.
(5) 해집단의 품질을 높이기 위해 작업을 일정 횟수 동안 반복한다.
(6) 위의 과정을 비디오가 끝날 때까지 반복하고 마지막으로 남는 하이퍼에지에 의해 하이퍼네트워크가 생성된다.

문장 생성 알고리즘은 다음과 같다.

(1) 특정 한 단어 *w*를 선정한다.
(2) SRILM에서 만들어진 언어모델을 이용해 *w*의 오른쪽에 위치할 단어를 확률적으로 생성한다.
(3) 문장 끝 기호인 </s>가 나오거나 일정 단어 개수를 채울 때까지 반복 생성한다.
(4) SRILM에서 만들어진 언어모델을 이용해 w의 왼쪽에 위치할 단어를 확률적으로 생성한다.
(5) 문장 처음 기호인 <s>가 나오거나 일정 단어 개수를 채울 때까지 반복 생성한다.

3. 실 험 결 과

멀티 모달 개념 망은 517분 길이의 'Maisy' 만화 비디오를 학습한 것을 사용했다. 언어 모델은 SRILM을 사용하여 'Maisy' 만화 비디오의 에피소드 1부터 6까지 모든 자막을 코퍼스로 학습했다. 학습한 멀티 모달 개념 망 중 mouse와 rabbit의 개념 망은 그림 2와 같았다.

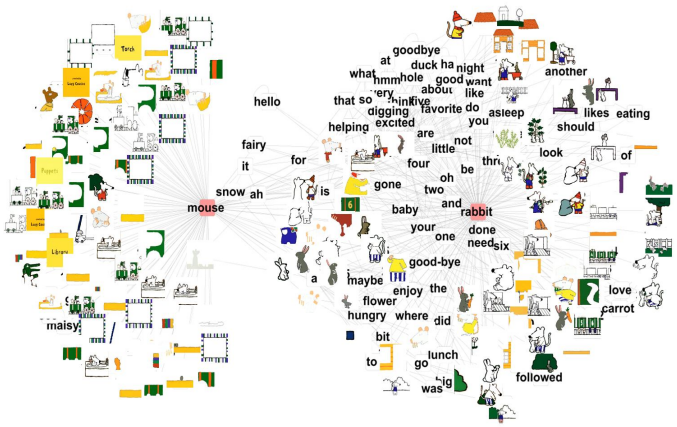


그림 2. 'Maisy' 비디오에서 학습된 mouse와 rabbit의 개념 망

그리고 특정 화면 이미지를 개념 망을 사용해 모달리티 교차 질의 확장을 한 결과는 그림 3과 같았다.

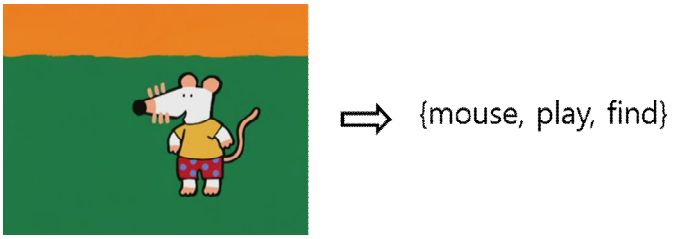


그림 3. 이미지에 대한 모달리티 교차 질의 확장 결과

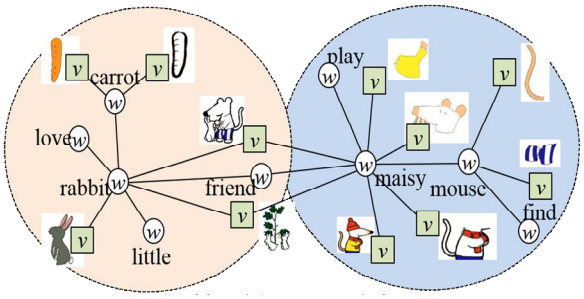


그림 4. 개념을 설명할 수 있는 이미지 패치와 단어들로 구성된 개념 망

질의 변경을 하기 위해 개념 망에서 개념을 설명할 수 있는 단어를 선정하는 작업이 필요했는데 이 때 개념 망은 그림 4와 같이 임의로 표현될 수 있었다.
또한 언어 모델을 사용해서 그림 3의 단어들을 포함한 문장을 생성해본 결과는 표1과 같았다. 문장은 w를 mouse로 선정한 뒤 생성했고 생성된 문장들 중 모달리티 교차 질의 확장 결과로 나온 단어를 모두 포함하는 문장들만 선택했다.

표1. 언어 모델을 사용해서 생성한 이미지 설명문

생성된 문장
play left. yellow. find these shorts rinse fairy mouse
[Mumbling] Ba-bye! [Tooting] Row, row, row Ahoy home, find? Feathers mouse Go! Go! to play.
toilet? find? asleep. Your Yellow She present! Dolphin hide-and-seek? back, feather fairy mouse It's play Cyril
everybody? find? playroom? Eh? soccer? nice reeds. fairy mouse be?
walking stick... find Yes, hurrey! play tennis, swim next. Munching, sink. fairy mouse
Hmm. Maybe your arms playground. find starting

문장들은 문법을 고려하지 않고 생성되었기 때문에 완전한 문장은 아니었다. 그리고 언어 모델이 비디오의 전체 자막을 대상으로 학습되었기 때문에 이미지와 관련이 없는 단어들이 생성된 경우도 있었다. 하지만 문법을 고려한 문장 생성과 언어 모델에서 더 적합한 단어를 선정하는 문제가 해결이 된다면 본래 화면 이미지의 해당 자막인 'Can you think of a hopping game to play, Maisy?'보다 더 적절한 이미지 설명문을 생성할 가능성이 보였다.
또한 개념 망에서 개념을 설명하는 단어를 선택할 때 그 기준을 마련하는 점도 필요했다.

4. 결론 및 향후 연구 방향

본 논문에서는 멀티모달 하이퍼네트워크 모델의 학습을 통해 'Maisy' 만화 비디오로부터 멀티모달 개념 망을 생성하였고 특정 화면 이미지를 설명할 수 있는 문장을 생성하는 방법을 제시하였다. 본 연구에서 사용된 유아용 만화 비디오는 실세계에서 쉽게 구할 수 있을 뿐 아니라 이미지 프로세싱을 최소로 줄이고 문제의 복잡도를 줄일 수 있기 때문에 본 연구의 테스트베드로 적합하다. 생성된 문장은 문법이 맞지 않다는 점과 이미지와 관련이 없는 단어가 포함되는 경우가 존재한다는 측면에서 개선의 여지가 있다. 하지만 이 문제가 해결된다면 적절한 이미지 설명문이 될 수 있는 가능성을 확인할 수 있었다. 향후 모달리티 교차 질의 확장을 할 때 개념 망에서 적절한 단어를 선정하는 기준을 마련하는 연구가 진행되어야 할 것이다.

5. 감사의 글

이 논문은 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734, Vldeome 및, NRF-2013M3B5A2035921, HyperIntelligence), 산업통상자원부의 SW컴퓨팅산업원천기술개발사업(10035348, mLife)에 의해 일부 지원되었음

6. 참 고 문 헌

[1] Fan Bu, Yu Hao, and Xiaoyan Zhu, Semantic relationship discovery with Wikipedia structure, in *Proc. of the 22nd International Joint Conference on Artificial Intelligence* (*IJCAI 2011*), pages 1770-177, 2011.

[2] C. Fellabaum, "WordNet," In R. Poli et al. (eds), *Theory and Applications of Ontology Applications*, pp. 231-243, Springer Science+Business Media, 2010.

[3] Jung-Woo Ha, Bum-Jin Lee, and Byoung-Tak Zhang, Text-to-Image retrieval based on incremental association via multimodal hypernetworks, in *Proc. of 2012 IEEE Conference on Systems, Man, and Cybernetics* (*IEEE SMC 2012*), pp. 3239-3244, 2012.

[4] Byoung-Tak Zhang, Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3), pp. 49-63, 2008

[5] Byoung-Tak Zhang, Jung-Woo Ha, and Myunggu Kang, Sparse population code models of word learning in concept drift, In *Proc. of Annual Meeting of the Cognitive Science Society* (*CogSci 2012*), pp. 1221-1226, 2012.

[6] A. Stolcke, SRILM - An Extensible Language Modeling Toolkit, in *Proc. International Conference on Spoken Language Processing (ICSLP 2002.)*, Vol. 2, pp. 901-904, 2002

[7] Jiri Matas, O. Chum, M. Urban, and T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, *Image and Vision Computing*, 22(10), pp. 761-767, 2004.

[8] David G. Lowe, Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60), pp. 91-110, 2004.

[9] Beom-Jin Lee, Jung-Woo Ha, Kyung-min Kim and Byoung-Tak Zhang, Evolutionary Concept Learning from Cartoon Videos by Multimodal Hypernetworks, in *Proc. of IEEE Congress on Evolutionary Computation* (*CEC2013*), (to appear)