멀티모달 개념계층모델을 이용한 만화비디오 컨텐츠 학습을 통한 등장인물 기반 비디오 자막 생성

김경민 ○, 하정우, 이범진, 장병탁

서울대학교 컴퓨터공학부

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

Character-based subtitle generation by learning of multimodal concept hierarchy from cartoon videos

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

School of Computer Science and Engineering, Seoul National University

요 약

기존 멀티모달 학습 기법의 대부분은 데이터에 포함된 컨텐츠 모델링을 통한 지식획득 보다는 이미지나 비디오 검색 및 태깅 등 구체적 문제 해결에 집중되어 있었다. 본 논문에서는 멀티모달 개념계층모델을 이용하여 만화 비디오로부터 컨텐츠를 학습하는 기법을 제안하고 학습된 모델로부터 등장인물의 특성을 고려한 자막을 생성하는 방법을 제시한다. 멀티모달 개념계층 모델은 개념변수층과 단어와 이미지 패치의 고차 패턴을 표현하는 멀티모달 하이퍼네트워크층으로 구성되며 이러한 모델구조를 통해 각각의 개념변수는 단어와 이미지패치 변수들의 확률분포로 표현된다. 제안하는 모델은 비디오의 자막과 화면 이미지로부터 등장 인물의 특성을 개념으로서 학습하며 이는 순차적 베이지안 학습으로 설명된다. 그리고 학습된 개념을 기반으로 텍스트 질의가 주어질 때 등장인물의 특성을 고려한 비디오 자막을 생성한다. 실험을 위해 총 268분 상영시간의 유아용 비디오 '뽀로로'로부터 등장인물들의 개념이 학습되고 학습된 모델로부터 각각의 등장인물의 특성을 고려한 자막 문장을 생성했으며 이를 기존의 멀티모달 학습모델과 비교했다. 실험결과는 멀티모달 개념계층모델은 다른 모델들에 비해 더 정확한 자막 문장이 생성됨을 보여준다. 또한 동일한 질의어에 대해서도 등장인물의 특성을 반영하는 다양한 문장이 생성됨을 확인하였다.

1. 서론.

스마트폰과 유튜브 등 IT의 발전을 통해 이미지, 동영상 데이터가 급격하게 증가함에 따라 멀티모달 데이터로부터 지식을 학습하는 기법에 대한 연구가 활발하게 진행되고 있다. 최근 deep learning 또는 비모수 베이지안 모델을 비롯한 다양한 멀티모달 학습 기법이 연구되어 왔으나[1, 2] 대부분은 데이터에 포함된 컨텐츠를 모델링하여 지식을 학습하기 보다는 이미지나 비디오 검색 및 태깅 등 구체적인 문제 해결에 집중되어 있었다. 본 논문에서는 지속적인 데이터의 증가에 따른 변화하는 개념을 효과적으로 학습할 수 있는 멀티모달 개념계층모델을 소개하고 유아용 만화 비디오로부터 컨텐츠를 모델링하는 기법을 제안한다. 멀티모달 개념계층모델(Multimodal concept hierarchy)은 이미지-텍스트 기반 개념 표현을 위해 Zhang에 의해 제안된 SPC (Sparse Population Coding) 모델[3]과 달리, 계층적 구조로 구성되며 하위층은 SPC와 같이 단어와 이미지 패치의 고차 패턴을 표현하는 하이퍼그래프(hypergraph) 구조[4]를 포함하고 상위층은 개념변수들로 구성되며 이 개념변수들은 하위층의 유사한 특성을 갖는 하이퍼에지(hyperedge)들로 구성된 부분그래프와 연결된다. 본 연구에서 개념변수는 등장 인물의 특성을 나타내며 만화 비디오로부터의 등장인물 개념 학습은 순차적 베이지안 추론으로 설명된다. 그리고 텍스트 질의가 주어질 때 학습된 개념을 바탕으로 멀티모달 추론을 통해 등장 인물의 특성을 고려한 비디오 자막을 생성한다. 실험을 위해 총 52개 에피소드 268분 상영시간의 유아용 만화 비디오 '뽀로로 시즌 3'가 사용되었다. '뽀로로'의 등장인물들의 개념을 학습시킨 후 모델로부터 등장인물의 특성을 고려한 비디오의 자막을 생성하고 이를 기존의 멀티모달 학습모델과 비교해본 결과 멀티모달 개념계층모델이 다른 모델들보다 더 정확한 자막 문장을 생성함을 확인했다. 또한 동일한 질의어에 대해서도 등장인물의 특성을 반영하는 다양한 문장을 생성함을 확인하였다.

2. 멀티모달 개념계층모델

멀티모달 개념계층 모델(Multimodal Concept Hierarchy: MuCH)은 계층구조로 표현되며 하위층은 하이퍼그래프 구조를 이용하여 단어와 이미지 패치의 고차 패턴을 표현하는 하이퍼에지 집합으로 구성된다. 상위층은 하이퍼에지의 부분집합과 연결되어 있는 개념변수들을

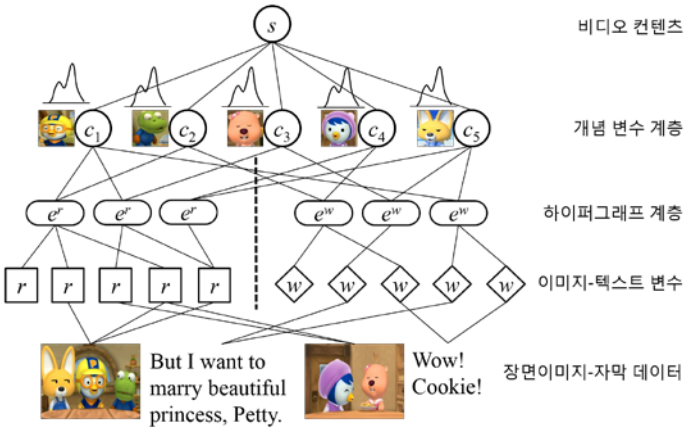


그림1. 멀티모달 개념계층모델의 구조

포함하며 본 연구에서는 하나의 개념 변수는 각 등장인물에 대응된다(그림 1). 비디오의 화면이미지와 자막 데이터로부터 이미지 패치들과, 단어들이 추출되고 추출된 이미지 패치와 단어는 그래프에서 각각 하나의 노드에 대응된다. 그러므로 하이퍼에지들은 단어와 이미지 패치의 고차 패턴을 표현할 수 있다[3]. 하이퍼에지를 구성하는 이미지 패치와 단어들의 패턴 연관성은 하이퍼에지의 가중치로 표현된다. 그리고 상위층의 개념변수들은 하위층에 존재하는 개념변수와 연관성이 큰 패치 및 단어들로 구성된 하이퍼에지 부분집합들과 연결되며 개념변수들은 하이퍼에지를 공유할 수 있다. 이러한 모델구조를 통해 등장인물들은 단어와 이미지패치 변수들의 확률분포로 표현된다.

이를 수식으로 정의하면 화면이미지-자막 데이터에서 추출된 이미지 패치와 단어는 각각 이진벡터인 $\mathbf{r}=(r_1,\ldots,r_N)$과 $\mathbf{w}=(w_1,\ldots,w_M)$으로 표현되고 등장인물 정보는 이진벡터 $\mathbf{c}=(c_1,\ldots,c_K)$로 나타내어진다. 이때, 모델의 파라미터 $\theta=(\mathbf{e}, \boldsymbol{\alpha})$와 등장인물 정보 $\mathbf{c}$가 주어졌을 때 이미지 패치-자막 쌍의 확률분포는 다음과 같다.

$$P(\mathbf{r},\mathbf{w}\,|\,\mathbf{c})=\sum_{\mathbf{e},\boldsymbol{\alpha}} P(\mathbf{r},\mathbf{w}\,|\,\mathbf{e},\boldsymbol{\alpha},\mathbf{c})P(\mathbf{e},\boldsymbol{\alpha}\,|\,\mathbf{c})\,, \qquad (1)$$

여기서 $\mathbf{e}$는 하이퍼에지의 집합을, $\boldsymbol{\alpha}$는 하이퍼에지 가중치 집합을 의미한다. 멀티모달 계층모델은 비디오에서 스토리가 진행됨에 따라 순차적으로 학습하게 되고 학습 단위는 에피소드 하나씩이다. 학습은 순차적 베이지안 추론을 통해 이뤄지고 식은 다음과 같다.

$$P_t(\mathbf{e},\boldsymbol{\alpha}\,|\,\mathbf{r},\mathbf{w},\mathbf{c})=\frac{P(\mathbf{r},\mathbf{w},|\,\mathbf{c},\mathbf{e},\boldsymbol{\alpha})P(\mathbf{c}\,|\,\mathbf{e},\boldsymbol{\alpha})P_{t-1}(\mathbf{e},\boldsymbol{\alpha})}{P(\mathbf{r},\mathbf{w},\mathbf{c})}, \qquad (2)$$

p_t는 t번째 에피소드에서 확률분포를 의미하고 t번째 에피소드가 들어왔을 때 prior 분포 $p_{t-1}(\theta)$는 posterior 분포를 계산하는데 사용이 된다. 여기서 계산된 posterior 분포 $p_t(\theta)$는 다음 단계에서 prior로 사용된다. 식 2는 또 다음과 같이 변형될 수 있다.

$$P_t(\mathbf{e},\boldsymbol{\alpha}\,|\,\mathbf{r},\mathbf{w},\mathbf{c})\propto\prod_{d=1}^{D_t}\left\{P(\mathbf{r}^{(d)},\mathbf{w}^{(d)}\,|\,\mathbf{c},\mathbf{e},\boldsymbol{\alpha})P(\mathbf{c}\,|\,\mathbf{e},\boldsymbol{\alpha})P_{t-1}(\mathbf{e},\boldsymbol{\alpha})\right\}, \qquad (3)$$

여기서 D_t는 t번째 에피소드의 데이터 크기이다. 학습은 log함수를 사용하여 log likelihood를 최대화하는 방식으로 이뤄진다.

$$\theta'=\arg\max_{\theta}\left[\left\{\sum_{d=1}^{D^t}\left(\log P(\mathbf{r}^{(d)},\mathbf{w}^{(d)}\,|\,\mathbf{c}^{(d)},\mathbf{e},\boldsymbol{\alpha})+\log P(\mathbf{c}^{(d)}\,|\,\mathbf{e},\boldsymbol{\alpha})\right)+D^t\log P_{t-1}(\mathbf{e},\boldsymbol{\alpha})\right\}\right]. \qquad (4)$$

첫번째 식은 등장인물정보와 모델의 파라미터가 주어졌을 때 단어와 이미지 패치 생성과 관련이 있고 두번째 식은 모델을 통해 등장인물의 동시등장 예측과 관련이 있다. 마지막 식은 지난 단계에서 학습한 모델을 반영한다.

3. 문장 생성 알고리즘

학습된 개념을 기반으로 텍스트 질의가 주어졌을 때 등장인물의 특성을 고려한 비디오 자막 생성은 아래 식과 같이 멀티모달 추론으로 설명된다.

$$P(\mathbf{w}\,|\,\mathbf{r},\mathbf{c},\mathbf{e},\mathbf{a})=\frac{P(\mathbf{w},\mathbf{c}\,|\,\mathbf{r},\mathbf{e},\boldsymbol{\alpha})}{P(\mathbf{c}\,|\,\mathbf{r},\mathbf{e},\boldsymbol{\alpha})}. \qquad (5)$$

위 식에서 $P(\mathbf{c}\mid\mathbf{r},\mathbf{e},\boldsymbol{\alpha})$는 등장인물과 연결된 부분하이퍼그래프로부터 계산된다. 본 연구에서는 문장생성을 위해 [5]에 제안된 방법을 이용하였으며 구체적인 알고리즘은 아래와 같이 기술된다.

단계1. 키워드 $X_q=\{x_q\}$가 주어졌을 때 x_q를 포함하고 등장인물 c와 연결된 하이퍼에지 부분집합 $M=\{e_1, e_2, \ldots, e_m\}$을 찾는다.

단계2. M에서 룰렛휠 선택방식을 이용해 $e_q=\{x_{q-1},x_q,x_{q+1}\}$인 하이퍼에지를 선택한다.

단계3. $X_q=\{x_{q-1}\}$로 설정하고 단계 1,2를 반복한다. 그 결과 하이퍼에지 $e_l=\{x_{q-2},x_{q-1},x_q\}$가 선택된다.

단계4. $X_q=\{x_{q+1}\}$로 설정하고 단계 1,2를 반복한다. 그 결과 하이퍼에지 $e_r=\{x_q,x_{q+1},x_{q+2}\}$가 선택된다.

단계5. e_l와 e_r를 연결하여 부분 문장 $s_q=\{x_{q-2}, x_{q-1},x_q,x_{q+1},x_{q+2}\}$를 생성한다.

4. 실 험 결 과

4.1 데이터 전처리

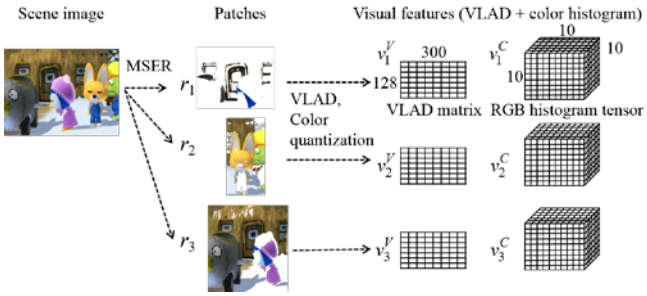
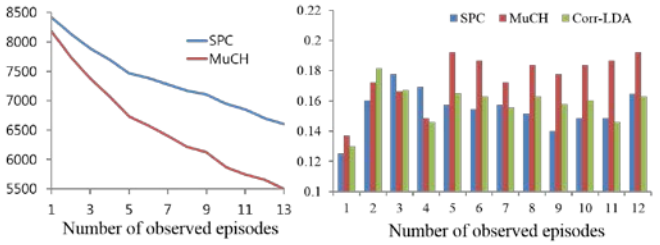


그림2. 이미지 패치 전처리 과정



(a) (b)

그림3. 스토리가 진행됨에 따른 모델 성능

비디오 데이터는 약 5000개의 화면이미지-자막 쌍으로 변환하였다. 그리고 그림 2와 같이 화면이미지로부터 MSER(Maximally Stable External Regions)를 사용하여 이미지 패치를 추출한 뒤 각각의 패치를 SIFT(Scale-invariant feature transform)을 사용하여 인코딩한 뒤 다시 VLAD(Vector of Locally Aggregated Descriptor)[6]로 128 x k 차원의 행렬로 인코딩하였다. 이 때 k는 SIFT 클러스터의 크기로 300을 사용하였다. 또한 픽셀의 RGB 값을 10 단계로 양자화 하여 각 이미지 패치마다 10x10x10 크기의 RGB 히스토그램을 계산하였다. 자막은 이진벡터로 표현이 되었다.

4.2 문장 생성

그림 3은 학습이 완료된 멀티모달 개념계층모델과 다른 멀티모달 모델들이 자막생성한 뒤 성능을 비교하고 있다. (a)는 likelihood를 나타내고 (b)는 아직 보여지지 않은 에피소드 13의 이미지가 질의로 주어졌을 때 생성한 단어의 precision을 나타낸다. 멀티모달 개념계층모델은 SPC모델이나 Corr-LDA모델[7]보다 더 좋은 성능을 나타냄을 확인할 수 있었다. 표1은 동일한 질의어에 대해서 등장인물 정보가 주어졌을 때 생성된 자막을 보여주고 있다. 표1로부터 동일한 질의어에 대해서도 등장인물에 따라 서로 다른 자막문장이 생성됨을 알 수 있으며 이는 제안하는 모델의 개념학습을 통해 등장인물의 정보가 문장 생성에 반영됨을 의미한다.

5. 결론 및 향후 연구 방향

본 논문에서는 멀티모달 개념계층모델을 이용하여 만화 비디오로부터 컨텐츠를 학습하고 학습된 모델로부터 등장인물 정보를 반영한 비디오 자막을 생성해보았다. 문장 생성 결과 다른 모델보다 정확하고 등장인물의 정보를 더 반영한 문장을 생성할 수 있었다. 향후 연구로는 멀티모달 개념계층모델을 통해 갤럭시기어나 구글글라스 등의 휴대용 기기에서 얻어지는 데이터로부터 지식습득을 해보는 것이다

6. 감사의 글

이 논문은 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734, Videome), 산업통상자원부의 SW컴퓨팅산업원천기술 개발 사업(10035348, mLife)에 의해 일부 지원되었음

표1. 등장 인물 기반 비디오 자막 생성 결과

질의어	등장인물(Tongtong) 정보 반영
 clock, I have made another potion come and try it (질의어: i, try)	as i don't have the right magic potion come and try it was nice
	ah, finished i finally made another potion come and try it we'll all alone?
	등장인물 정보 미반영
	yes.. come and try it before i forget
	as i try to get the doll keeps saying he's alright

7. 참 고 문 헌

[1] N. Srivastava, and R. Salakutdinov, Multimodal Learning with Deep Boltzmann Machines, *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pp. 2231—2239, 2012

[2] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, Learning Systems of Concepts with an Infinite Relational Model. *In Proceedings of the 21st Conference on Artificial Intelligence (AAAI 2006)*, 381-388, 2006

[3] B. T. Zhang, J. W. Ha, and M. G. 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.

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

[5] J. H. Lee, S. H. Lee, W. H. Chung, E. S. Lee, T. H. Park, R. Deaton, and B.-T. Zhang, A DNA assembly model of sentence generation, *BioSystems*, 106:51-56, 2011.

[6] H. Jegous, M. Douze, C. Schmid, and P. Perez, Aggregating Local Descriptors into a Compact Image Representation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*. 3304-3311, 2010

[7] H. Xiao, and T. Stibor, Toward Artificial Synesthesia: Linking Images and Sounds via Words. *NIPS Workshop on Machine Learning for Next Generation Computer Vision Challenges*, 2010