만화비디오를 학습하는 심층 하이퍼네트워크의 최적 시각 특징 탐색

강우영 ○ 김경민 남장군 장병탁
서울대학교 컴퓨터공학부
{wykang, kmkim, cjnan, btzhang}@bi.snu.ac.kr

Visual Features Analysis for Deep Hypernetworks Learning Cartoon Videos

Woo-Young Kang, Kyung-Min Kim, Chan-Jun Nan, Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

최근 인간수준의 인공지능을 실현하기 위해 많은 연구들이 진행되고 있으며, 그 중 멀티모달 학습 기법 연구는 인간이 다양한 감각으로부터 오는 상이한 패턴의 데이터를 학습한다는 점에서 매우 중요하다. 심층 하이퍼네트워크는 연상 기억 메모리 방식과 sparse population coding 기법을 사용하여 다양한 감각으로부터 얻어지는 멀티모달 데이터를 학습할 수 있으며 지금까지 만화 비디오 지식 자동 구축을 이용한 이미지-문장 상호 변환, 동영상의 이미지 자동 태깅 등에 사용되어왔다. 언급한 예에서 심층 하이퍼네트워크는 이미지와 텍스트를 입력 데이터로 받으며 컴퓨터 비전 분야의 기술 발전에 따른 다양한 이미지 처리 기법들을 사용하여 데이터를 전처리 하였다. 하지만 아직까지 심층 하이퍼네트워크의 최적 성능을 위한 이미지 처리 기법에 대한 연구가 제대로 이루어지지 않고 있다. 본 논문에서는 심층 하이퍼네트워크를 위한 최적 시각 특징을 찾기 위해 MSER 과 R-CNN 이미지 전처리 기법에 따른 모델의 성능을 비교한다. 실험을 위해 MSER 과 R-CNN 으로부터 이미지 특징들을 추출하였고, 각각 서로 다른 심층 하이퍼네트워크의 시각 입력 데이터로 사용한 뒤, 그에 따른 모델의 추론 성능을 비교하였다. 데이터로 에피소드 52 편 300 분 분량의 어린이 만화영화 '뽀로로'를 사용하였고, 하나의 장면을 주고 대사를 생성한 뒤, precision/recall 로 모델의 성능을 분석하였다.

1. 서 론

인간의 사고 과정을 이해하고 그를 모방한 인간 수준의 인공지능 시스템을 만드는 문제는 현대 인공지능 분야의 큰 화두이며 매우 도전적인 과제이다. 이러한 인간 수준의 인공지능을 만들기 위해서는 인간처럼 학습하고 학습된 지식으로부터 새로운 정보를 추론하거나 생성해 내는 상상력이 있어야 한다. 또한, 인간의 학습은 유니모달한 데이터 보다는 멀티모달한 데이터로부터 이루어 지므로, 하나의 자극이 아닌 다양한 감각들로부터 얻어지는 멀티모달 데이터를 효과적으로 모델링 할 수 있어야 한다. 심층 하이퍼네트워크는 주어진 멀티모달한 데이터로부터 Markov Chain Monte Carlo(MCMC)기반의 학습기법을 통해 개념망을 구축하여 지식을 저장하고 저장된 지식들로부터 역으로 유사성에 기반한 연상작용을 통해 새로운 정보를 추론하는 일종의 상상을 할 수 있다[1]. 멀티모달 데이터의 한 예로는 시각-언어 쌍을 들 수 있는데, 이미지-대사 쌍을 입력으로 하는 심층 하이퍼네트워크 모델은 구축된 개념구조로부터 추론을 통해 의미적으로 유사한 이미지-자막을 상호교차 생성시킬 수 있다. 본 연구에서는 52 편 분량의 뽀로로 만화영화에서 자막이 나오는 프레임들을 선별하여, 이미지-문장 쌍을 만든 후, 각 이미지 프레임들로부터 두 가지 객체 탐지 기법인 Maximally Stable External Regions(MSER) 기법[2]과 Regions with Convolutional Neural Networks features(R-CNN) 기법[3]을 사용하여 각각 시각특징을 추출해 낸다. 이후 동일한 언어특징과 두 가지 다른 시각특징을 모델의 입력으로 사용하여 개념망을 각각 구축한 뒤 구축한 개념구조를 사용하여 이미지를 주었을 때 각 모델이 자막을 생성하는 자막생성 실험을 수행한다. 실험을 통해 두 모델의 자막생성 성능에 어떠한 차이가 있는지 분석하여 심층 하이퍼네트워크에 최적화된 시각특징에 대해 논의한다.

2. 심층 하이퍼네트워크

하이퍼네트워크는 최초 DNA 컴퓨팅 기반으로 알고리즘이 고안되었다[4]. 이후 범용적인 컴퓨터 연산기반 하에 여러

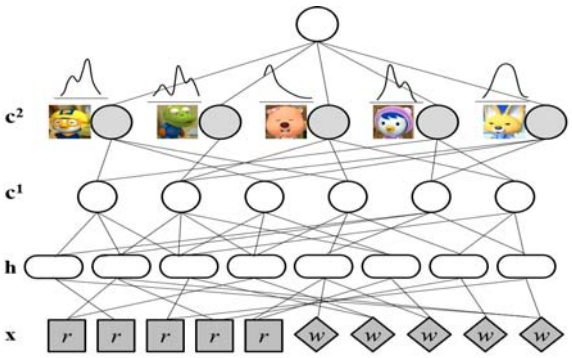


그림 1. 심층 하이퍼네트워크 개념망의 예시

가지 응용들로 그 활용분야를 넓혀가고 있다[5][6]. 심층 하이퍼네트워크는 명시적인 개념 층 들을 사용하여 추상적인 개념들을 표현할 수 있다. 심층 하이퍼네트워크의 층 사이 연결은 성긴 표현(sparse representation)을 가지며 각 은닉층 노드들의 수와 연결이 동적으로 학습된다. 이러한 성김 모델은 인간의 뇌에서 발견할 수 있는 성기며 계층적인 모듈성을 모사한다고 볼 수 있다. 뽀로로 비디오로부터 추출된 각 프레임들을 시각특징과 언어특징으로 나누고, 이들을 사용하여 심층 하이퍼네트워크 모델을 구성하면 그림 1 과 같은 모델이 만들어 진다.

3. 객체 분리

최초 선별된 이미지 프레임 안에는 관심 있는 주요 등장인물 외에 배경화면 및 비 관심 객체 등 많은 객체들이 있다. 이 중에서 관심 있는 주요 등장인물 이미지 패치를 추출하기 위해서는 별도의 객체탐지 기법을 사용하여야 한다. 객체를 탐지하는 알고리즘으로는 다양한 방법이 존재하지만 본 실험에서는 대표적인 두 기법인 MSER 과 R-CNN 기법을 사용한다.

3.1 MSER

MSER 기법은 하나의 이미지로부터 특정 임계 값들의 변화에 대해 상관성이 있는 지역들을 extremal region 이라 불리는 연결성분들로 묶어 객체들을 추출해 내는 알고리즘이며[2], 처리속도를 끌어올리기 위해 component tree 구조를 사용하는 방법도 제시되었다[7]. Extremal regions R_i는 수식 (1)과 같이 정의된다.

$$\forall p \in R_i, \forall q \in boundary(R_i) \rightarrow I_{in}(p) \geq I_{in}(q) \qquad (1)$$

이후 찾아진 extremal regions 으로부터 MSERs 는 다음과 같이 정의된다.

$$Q(R_i^g) = (|R_j^{g-\Delta}| - |R_j^{g+\Delta}|)/|R_i^g| \qquad (2)$$

식(2)에서 |·|는 cardinality 를 나타내며, R_i^g는 그레이 레벨 g 로 thresholding 된 extremal region 을 나타내고, Δ는 stability range 파라미터이다. 이렇게 찾아진 extremal region 은 이미지 좌표계의 연속적인 일대일 변화와 이미지 밝기의 선형적인 변화에 강건한 중요한 두 특징이 있다.

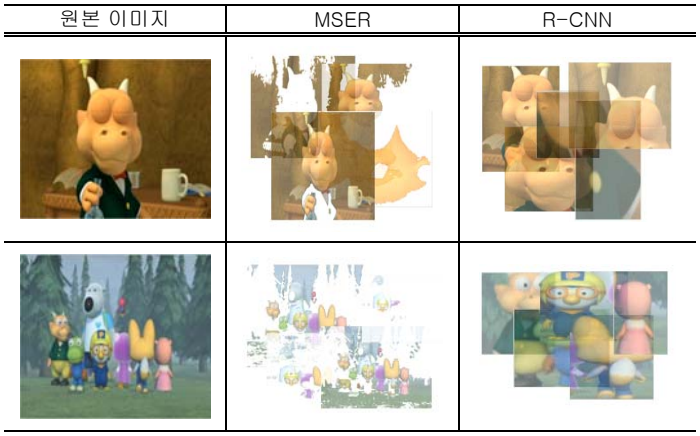


그림 2. MSER 과 R-CNN 에 의해 추출된 이미지패치

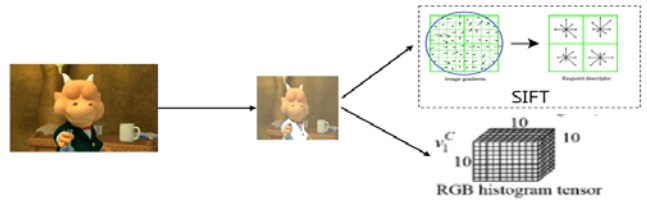
3.2 R-CNN

R-CNN 은 최근 객체분류에서 우수한 성능을 보이는 CNN 기법[8]을 약간 변형한 것으로, 모델의 입력으로 들어가는 데이터가 원본 이미지가 아닌 selective search[9]를 통해 영상 내에서 각 객체를 포함하는 sub-region 을 추출한 뒤 CNN 의 입력으로 한다. 이후 추출된 특징벡터를 SVM 을 통해 분류한다. 하지만 본 실험에서는 분류 전 단계인 특징추출단계 까지만 사용하였다. 두 방법으로 추출된 각 이미지 패치는 그림 2 와 같다.

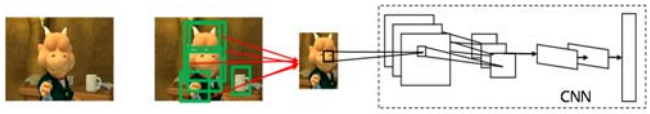
4. 실험 결과

4.1 데이터 전처리

실험에 사용된 데이터는 어린이 만화영화인 뽀로로의 에피소드 총 183 편 중 52 편을 사용하였다. 먼저 한 장면에서 대사를 추출해 단어 단위로 분리한 뒤 word-2-vec 기법[10]을 통해 단어를 실수벡터로 표현하여 언어특징



(a) MSER 을 통한 시각특징 추출



(b) R-CNN 을 통한 시각특징 추출

그림 3. 시각특징 표현

표 1. MSER 과 R-CNN 에 따른 문장생성 분석결과

	Measure	MSER	R-CNN
Image 1	Precision	0.1664	0.1587
	Recall	0.2508	0.2608
	F-score	0.2001	0.1973
Image 2	Precision	0.1288	0.1243
	Recall	0.2430	0.2370
	F-score	0.1687	0.1631

을 만든다. 이후 MSER 로 추출된 각 이미지 패치는 두 종류의 벡터로 표현된다. 먼저 하나는 각 RGB 값들을 10 개의 구간으로 나누어 1000 차원 히스토그램 벡터로 나타낸 RGB 컬러 히스토그램 벡터이고, 다른 하나는 SIFT[11] 히스토그램 벡터로 SIFT 로부터 추출한 특징 벡터들을 클러스터링 한 뒤 히스토그램 벡터로 나타낸 것이다. R-CNN 으로부터 추출된 각 이미지 패치 역시 두 개의 벡터로 표현되는데 하나는 앞서 말한 것과 같은 RGB 컬러 히스토그램 벡터이고 다른 하나는 R-CNN 으로부터 추출된 4096 차원의 CNN 특징벡터이다.

4.2 문장 생성

실험은 구성된 2 개의 심층 하이퍼네트워크 모델에 대하여 장면을 주고 그에 상응하는 대사를 생성해 내는 문장생성 실험을 수행하였다. 이를 위해 두 장의 이미지 프레임을 임의로 선별하였다. 이후, 선별된 두 이미지에 대하여 MSER 기반과 R-CNN 기반의 심층 하이퍼네트워크 모델이 생성해낸 문장이 정답 문장의 내용을 얼마나 잘 포함하는지를 precision/recall 기법을 사용하여 분석하였다. 실험 결과 문장 생성 성능에서는 MSER 을 사용한 방법이 precision 이나 recall 성능이 조금 앞섰지만, F-score 의 경우 큰 차이가 없었다. 하지만 MSER 의 경우 SIFT 로 추출된 시각특징의 경우 300 차원 특징을 사용한다. 이는 같은 크기의 컬러 히스토그램 벡터를 사용하지만, R-CNN 의 4096 차원 특징공간에 비해 크기가 매우 작다. 따라서 MSER 방식으로 추출된 시각특징이 본 실험의 경우 더욱 좋은 표현력을 가진다고 볼 수 있다.

5. 결 론

본 논문에서는 기존에 개발된 심층 하이퍼네트워크의 최적화된 시각 특징을 탐색하기 위한 실험을 수행하였다. 이를 위해 시각특징을 제외한 다른 변수들은 동일하게 유지한 채 서로 다른 두 가지 방법인 MSER 과 R-CNN 을 사용하여 각각 심층 하이퍼네트워크를 구성하였다. 이후 이미지 프레임이 주어졌을 때 문장을 생성해낸 뒤 생성된 문장이 정답문장의 내용을 얼마나 잘 포함하고 있는지를 precision/recall 기법을 통하여 분석하는 실험을 수행하였다. 그 결과 문장생성 성능은 MSER 기법을 사용한 심층 하이퍼네트워크 모델이 조금 우수하였다. 하지만 실험에 사용한 데이터가 비교적 단순하고 객체탐지가 쉬운 만화영화였다는 점을 고려해보면, 실제세계의 데이터를 넣었을 경우 일반적으로 R-CNN 이 더욱 좋은 객체 탐지 성능을 보이므로 더 좋은 문장생성 결과를 보여줄 수도 있을 것이다. 따라서 앞으로 실제 생활 데이터를 가지고 실험해보는 연구를 진행해보는 것이 향후 연구과제가 될 수 있을 것이다.

감사의 글

이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(R0126-15-1072-SW 스타랩, 10044009-HRI.MESSI)을 받아 수행된 연구임.

참고문헌

[1] J. W. Ha, K. M. Kim, and B. T. Zhang, “Automated Construction of Visual-Linguistic Knowledge via Concept Learning from Cartoon Videos”, *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, pp. 522-528, 2015.

[2] J. Matas, O. Chum, M. Urban and T. Pajdla, “Robust Wide Baseline Stereo from Maximally Stable External Regions”, *Image and Vision Computing* 22(10):761-767, 2004

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014

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

[5] K. M. Kim, J. W. Ha, B. T. Zhang, “Deep Hierarchical Networks for Learning Visually-Grounded Linguistic Concepts from Cartoon Videos”, *the Korean Institute of Information Scientists and Engineers (KIISE 2014)*, pp.518-520, 2014.

[6] K. M. Kim, J. W. Ha, C. J. Nan and B. T. Zhang, “Learning and Inference in Cartoon Videos using Deep Hypernetworks Concept Structure”, *Korea Computer Congress (KCC 2015)*, pp.814-816, 2015.

[7] M. Donoser and H. Bischof, “Efficient Maximally Stable Extremal Region (MSER) Tracking”, *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, 2006

[8] A. Krizhevsky, I. Sutskever, G. Hinton, “ImagNet Classification with Deep Convolutional Neural Networks”, *In Proceedings of Advances in Neural Information Processing Systems (NIPS 2012)*, 2012.

[9] J. R. R. Uijlings, K. E. A Van de Sande, T. Gevers and A. Smeulders, “Selective Search for Object Recognition”, *IJCV*, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representation of Words and Phrases and Their Compositionality”, *In Proceedings of Advances in Neural Information Processing Systems (NIPS 2013)*, 2013.

[11] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”. *International Journal of Computer Vision*, 2(60), pp. 91-110, 2004.