텍스트 정보와 시각 특징 정보를 이용한 효과적인 웹 이미지 캡션 추출 방법

황지익[O] 박주현 낭종호
서강대학교 컴퓨터학과
ziegh@mlneptune.sogang.ac.kr[O], {parkjh, jhnang}@sogang.ac.kr

An Efficient Web Image Caption Extraction Method based on Textual and Visual Information

Jiik Hwang[0], Joohyoun Park, Jongho Nang
Dept. of Computer Science, Sogang University

요 약

기존의 웹 이미지 검색 시스템들은 웹 페이지에 포함된 텍스트들의 출현빈도, 태그유형 등을 고려해 각 키워드들의 중요도를 평가하고 이를 이용해 이미지의 캡션을 결정한다. 하지만 텍스트 정보만으로 캡션을 결정할 경우, 키워드와 이미지 사이의 관련성을 평가할 수 없어 부적절한 캡션의 배제가 어렵고, 사람의 인지와 맞지 않는 캡션이 추출되는 문제점이 있다. 본 논문에서는 기존의 웹 이미지 마이닝 방법을 통해 웹 페이지로부터 캡션 후보 키워드를 추출하고, 자동 이미지 주석 방법을 통해 이미지의 개념 부류 키워드를 결정한 후, 두 종류의 키워드를 결합하여 캡션을 선택한다. 가능한 결합 방법으로는 키워드 병합 방법, 공통 키워드 추출 방법, 개념 부류 필터링 방법, 캡션 후보 필터링 방법 등이 있다. 실험에 의하면 키워드 병합 방법은 높은 재현율을 가져 이미지에 대한 다양한 주석이 가능하고 공통 키워드 추출 방법과 개념 부류 키워드 필터링 방법은 정확률이 높아 이미지에 대한 정확한 기술이 가능하다. 특히, 캡션 후보 키워드 필터링 방법은 기존의 방법에 비해 우수한 재현율과 정확률을 가지므로 기존의 방법에 비해 적은 개수의 캡션으로도 이미지를 정확하게 기술할 수 있으며 일반적인 웹 이미지 검색 시스템에 적용할 경우 효과적인 방법이다.

1. 서 론

인터넷 기술의 발달로 인하여 많은 사용자들이 WWW을 이용하게 되고 그에 따라 웹은 막대한 량의 이미지 데이터를 포함하는 멀티미디어 데이터 아카이브로서 활용할 수 있게 되었다. 이러한 대규모의 정보 집합체인 WWW에서 사용자가 원하는 이미지 데이터를 찾기란 결코 쉬운 일이 아니다. 따라서 WWW에서 사용자가 원하는 이미지 데이터를 정확하고 신속하게 찾아주는 정보 검색에 대한 연구가 널리 진행되어 오고 있다. 검색 속도와 Semantic Gap[11]등의 문제로 인하여 현재의 웹 이미지 검색은 웹 문서에 포함되어 있는 이미지 주변의 텍스트 정보를 이미지와 연결된 개념 정보로 간주하여 검색을 수행한다. 따라서 정확한 개념의 추출은 검색의 속도와 정확도의 향상을 의미하게 된다. 하지만 현재까지 연구된 웹 이미지 마이닝에 관련된 연구들[6,7,8,9,10]이 웹 페이지의 텍스트 정보에 대해서 출현빈도수와 태그 위치만을 고려하기 때문에 어떤 텍스트가 이미지에 포함된 개념과 높은 연관성을 갖는지 알 수 없다. 따라서 연관성이 낮은 캡션을 제거하지 못하여 추출되는 캡션의 개수가 많고 이미지에 포함된 개념과의 관련성은 낮다는 문제점을 가지고 있다.

이미지에 포함된 개념을 결정하는 또 다른 접근방법으로는 자동 이미지 주석 방법[1,2,3,4]이 있다. 이 방법에서는 미리 정해진 개념 부류에 속하는 이미지들로부터 시각적 특징 벡터를 추출하고 이를 통계적으로 분석하여 이미지의 저급 수준 정보와 고급 수준 정보 사이의 연관성을 학습하며 이를 이용해 주어진 이미지가 표현하고 있는 개념을 결정한다. 하지만, 이미지의 시각적 특징만으로 이미지가 표현하고 있는 개념을 정확하게 결정하기 어렵고 미리 이미지의 개념 부류를 정해야 한다는 점과 결정된 개념이 미리 정해놓은 이미지의 개념 부류에 의해 제한된다는 문제점을 가진다.

본 논문에서는 기존의 웹 이미지 마이닝 방법에 자동 이미지 주석 방법을 적용하여 보다 적은 개수의 캡션을 사용하면서 높은 정확도를 가지는 캡션 추출 방법을 제안하며 실험을 통해 성능을 분석하고 평가한다. 이를 위해서, 먼저 학습 단계에서 웹 페이지에 존재하는 이미지와 그 주변 텍스트를 사람이 직접 보고 캡션을 결정하여 학습 데이터를 구축한 후, 자동 이미지 주석 방법을 이용해 웹 이미지의 개념 부류와 시각적 특징 사이의 관련성을 모델링한다. 그 후, 웹 이미지 마이닝 방법을 이용하여 주어진 웹 이미지의 캡션 후보 키워드를 얻고 자동 이미지 주석 방법을 이용하여 이미지의 개념 부류 키워드를 얻어, 두 결과를 조합하여 최종 캡션으로 선택한다. 개념 부류 키워드와 캡션 후보 키워드를 결합하기 위해서 키워드 병합, 공통 키워드 추출, 개념부류 키워드 필터링, 캡션후보 키워드 필터링 등의 방법을 사용한다. 실험에 의하면 제안한 방법들은 기존의 방법에 비해 최대 3배 정도 높은 정확도를 가지게 됨을 알 수 있었다.

2. 기존의 연구

2.1 웹 이미지 검색을 위한 캡션 추출에 대한 기존의 연구

웹 페이지 상에서 이미지는 주변에 자신의 내용과 관련된 텍스트를 가지고 있다. 일반적으로 웹 페이지에 포함된 이미지와 텍스트들은 동일한 개념을 설명하고 있으므로 이미지에 표현된 개념과 관련이 높은 키워드들은 그렇지 않은 키워드들에 비해 자주 나타나게 된다. 따라서 기존의 웹 이미지 검색 시스템들은 이미지의 URL, 웹 페이지의 타이틀 및 이미지의 주변 텍스트 등을 그대로 키워드 캡션으로 사용하거나 이들에 대해 태그 위치 또는 출현 빈도수(TF)에 따른 가중치를 부여하여 사용한다. ImageRover[6]는 이미지를 포함하고 있는 웹 페이지의 HTML 태그로부터 키워드를

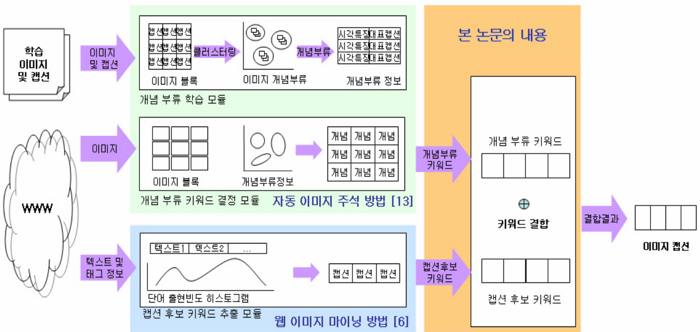
추출하는데 각 단어들이 위치한 HTML 태그에 따라 서로 다른 가중치를 적용한다. 즉, ALT 텍스트, 페이지 타이틀, 헤딩 및 볼드, 이탤릭 태그 등과 같이 이미지나 웹 페이지의 내용을 직접적으로 표현하고 있거나 시각적으로 강조된 텍스트와 이미지로부터 가까운 곳에 위치한 단어에 좀 더 높은 가중치를 부여한다. WebSeek[7], WebSeer[8], AMORE[9], MARIE[10]도 Image Rover와 거의 같은 방법을 사용하여 이미지 캡션을 추출한다.

2.2 자동 이미지 주석 방법

자동 이미지 주석 방법은 이미지에 포함된 의미를 나타내는 개념을 미리 정의하고 주어진 이미지에 이들로 주석을 다는 방법이다. 최근의 방법들은 일반적으로 서브 영역으로의 이미지 분할, 각 서브 영역의 내용 정보 표현, 그리고 내용 정보와 개념 정보간의 연결의 공통된 3가지 단계를 통해 이미지의 주석을 결정한다[1]. 마지막 과정에서는 HMM[2], Co-Occurrence 모델[3], SVM[4] 등과 같은 여러 종류의 통계적 학습 모델들이 사용되며 이런 모델을 통해 이미지 영역과 개념 사이의 관련성 확률을 유도한다.

3. 텍스트와 시각 특징을 이용한 캡션 추출 방법

본 논문에서는 캡션 추출의 정확성을 높이기 위해서 <그림 1>과 같이 기존의 HTML을 분석하여 캡션을 추출하는 방법[6]과 자동 이미지 자동 주석 방법[1]을 함께 사용하여 이미지에 대한 캡션을 선택한다.



<그림 1> 제안하는 이미지 캡션 추출 방법의 개관

즉, 두 방법을 함께 사용하여 키워드의 출현 빈도수에 의한 캡션 결정의 부정확성과 시각적 특징에 의한 개념 부류 결정의 부정확성을 보완할 수 있도록 두 종류의 키워드들을 적절히 평가하고 결합하여 최종 캡션을 선별한다. 이 때 가능한 방법은 키워드 병합, 공통 키워드 추출, 개념 부류 키워드 필터링, 그리고 캡션 후보 키워드 필터링의 4가지 방법을 생각해 볼 수 있다.

● 키워드 병합을 통한 캡션 선택 (방법1)

HTML을 분석하여 얻은 캡션 후보 키워드 집합과 이미지 자동 주석 방법을 사용하여 얻은 개념 부류 키워드 집합을 모두 캡션으로 선택한다. 이미지에 대한 가장 제세한 캡션 주석이 가능하지만 이미지와 관련성이 낮은 키워드들을 걸러 내지 못하므로 캡션의 정확도가 낮아지게 되고 키워드들의 중요도를 평가하지 못하므로 추출되는 캡션의 개수를 조절할 수 없다는 문제점을 갖게 된다.

● 공통 키워드 추출을 통한 캡션 선택(방법2)

캡션 후보 키워드 집합과 개념 부류 키워드 집합에 모두 포함되어 있는 키워드를 최종 캡션으로 선택하는 방법이다. 매우 높은 정확성을 보여주지만 동시에 공통된 키워드가 존재하지 않거나 이미지를 잘 표현해주는 키워드가 선택되지 않을 확률 역시 높아지는 문제점을 가지게 된다.

● 개념 부류 키워드 필터링을 통한 캡션 선택(방법3)

개념 부류 키워드와 캡션 후보 키워드 사이의 의미적 유사도를 측정하고 유사도에 따라 중요도를 부여한 후, 높은 중요도를 가진 개념 부류 키워드를 캡션으로 선택하는 방법이다. 이 때, 개념 부류 키워드 결정이 부정확할 수 있으므로 키워드 사이의 유사도뿐만 아니라 웹 페이지의 텍스트 정보로부터 얻어진 캡션 후보 키워드의 가중치를 반영하여 최종 캡션의 중요도를 결정한다. 개념 부류 키워드와 캡션 후보 키워드와의 의미적 유사도 워드넷[5]을 이용하여 계산되며, 워드넷 단어 그래프 상에서 두 키워드가 위치하는 노드들 사이의 최단 경로 길이에 의해 결정된다. 최종적으로 선택될 캡션의 중요도 계산은 아래의 수식과 같이 계산된다.

$$W_{Total} = \alpha \cdot W_{Similarity} + (1-\alpha) \cdot W_{CCKeyword} \quad ,0 \le \alpha \le 1 \qquad \text{<식 1>}$$

여기서 $W_{Similarity}$는 워드넷에 기반한 두 키워드 사이의 의미적 유사도에 의한 가중치를 말하며 $W_{CCKeyword}$는 기존의 웹 이미지 마이닝 방법에 의해 계산된 캡션 후보 키워드의 가중치를 의미하고 α는 캡션의 중요도에서 키워드 사이의 의미적 유사도가 차지하는 비율을 뜻한다. 이 방법은 개념 부류 키워드와 캡션 후보 키워드 사이의 의미적 유사도를 측정하여 최종 캡션 선택에 반영하기 때문에 동일한 개념이 서로 다른 키워드로 기술된 경우에도 정확한 캡션 결정이 가능하다는 장점을 갖는다. 하지만, 시각적 특징에 의한 이미지의 개념 부류 결정이 정확하지 못할 수 있으며, 미리 정해놓은 개념 부류에 의해 이미지의 캡션이 제한되기 때문에 웹 페이지상의 텍스트가 캡션으로 선택될 가능성이 낮다는 문제점을 갖게 된다.

● 캡션 후보 키워드 필터링을 통한 캡션 선택(방법4)

이 방법에서는 캡션 후보 키워드와 개념 부류 키워드 사이의 의미적 유사도를 측정하고 유사도에 따라 중요도를 부여한 후, 높은 중요도를 가진 캡션 후보 키워드를 캡션으로 선택하게 된다. 방법 3에서와 마찬가지로, 개념 부류 키워드 결정이 부정확할 수 있으므로 키워드 사이의 유사도뿐만 아니라 웹 페이지의 텍스트 정보로부터 얻어진 캡션 후보 키워드의 가중치를 반영하여 최종 캡션의 중요도를 결정한다. 캡션 후보 키워드와 개념 부류 키워드 사이의 의미적 유사도는 워드넷 단어 그래프 상에서 두 키워드가 위치하는 노드들 사이의 최단 경로 길이에 의해 결정된다. 최종적으로 선택될 캡션의 중요도 계산은 아래의 수식과 같이 계산된다.

$$W_{Total} = \alpha \cdot W_{Similarity} + (1-\alpha) \cdot W_{CCKeyword} \quad ,0 \le \alpha \le 1 \qquad \text{<식 2>}$$

이 방법을 사용하면 최종 캡션을 웹 페이지에 포함된 캡션 후보 키워드들로부터 결정하므로 미리 정한 개념 부류에 의해 캡션이 제한되는 문제점을 피할 수 있으며, 이미지의 시각적 특징으로부터 어떤 키워드가 이미지의 개념과 관련성이 높은 지에 대한 판단 기준을 얻을 수 있으므로 관련성이 낮은 캡션의 필터링이 가능하다.

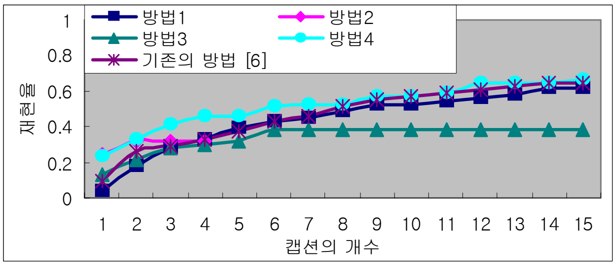
4. 실험 결과 및 분석

본 논문에서는 제안한 방법에 의한 캡션 추출 결과를 평가하고 기존의 방법과 비교를 위해서 사람에 의해 선택된 웹 이미지의 캡션과 제안한 방법 및 기존의 방법에 의해 선택된 캡션을 비교한다. 이에 더하여 제안한 방법이 다양한 개념 부류의 이미지를 포함하고 있는 WWW에 적용될 수 있는지를 알아보기 위해 시스템에 학습되어 있지 않은 개념 부류에 대한 캡션 추출 성능을 분석한다. 실험을 위해 사용한 데이터는 총 50개의 다양한 개념부류에 대해 WWW에서 수집한 580개의 이미지를 대상으로 한다. 수집된 이미지를 <표 1>과 같은 비율로 학습 이미지와 실험 이미지로 구성하였다.

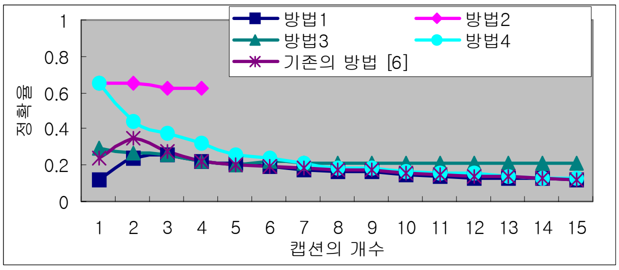
<표 1> 실험 데이터의 구성

	총 이미지 수	총 캡션 수	평균이미지 수	평균 캡션 수
학습 이미지	500개(86%)	1874개	10개/부류	3.7개/이미지
실험 이미지	80개(14%)	349개	1.6개/부류	4.3개/이미지

먼저, 제안한 방법의 일반적인 특성을 알아보기 위해 α의 값을 0.5로 설정한 상태에서 성능을 비교해보았다. <그림 2>은 각 방법을 사용하였을 때의 재현율과 정확률을 보여준다.



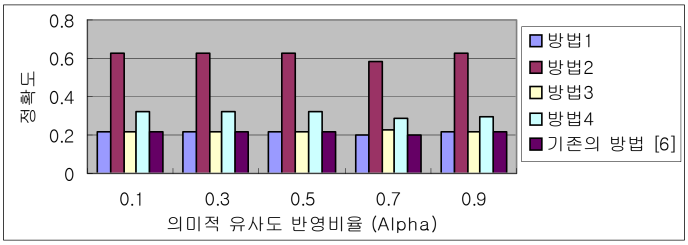
(a) 캡션 추출 방법의 재현률 비교 (α=0.5)



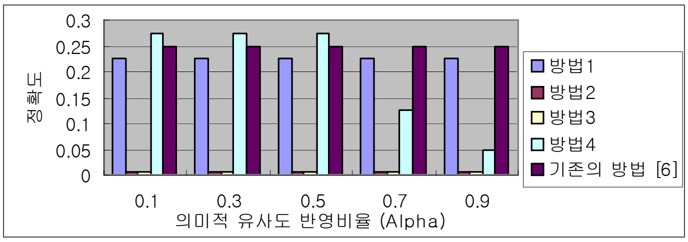
(b) 캡션 추출 방법의 정확률 비교 (α=0.5)

<그림 2> 캡션 추출 방법의 재현율과 정확률 비교

위의 표와 그래프로부터 방법1과 방법4가 다른 방법들에 비해 비교적 높은 재현율을 갖지만 정확률은 방법 2가 가장 좋은 성능을 보여주고 있음을 확인할 수 있다. 즉, 방법1은 재현율과 정확률이 기존의 방법과 큰 차이가 없으며 방법2와 3은 높은 정확률을 가지지만 추출되는 캡션의 개수가 적어 일반적인 웹 이미지 검색 시스템에 적용하기는 적합하지 않다. 방법4는 기존의 방법에 비해 재현율과 정확률모두 높게 측정되었다. 따라서 웹 이미지 검색 시스템에 방법 4를 적용하는 것이 기존의 방법에 비해 우수한 성능을 보일 것이다. 기존의 방법에 대한 성능 향상의 정도를 알아보기 위해 이미지 당 적절한 개수의 캡션을 추출하였을 때의 정확도를 비교해 본 결과는 <그림3>-(a)와 같다. 적절한 개수의 캡션은 사람에 의해 이미지의 캡션을 결정할 경우에 각 이미지에 할당된 평균 캡션의 개수인 3.7개를 기준으로 정해졌으며, 각 방법에 의해 4개의 캡션을 추출했을 때 정확한 캡션이 포함된 비율을 비교하였다. 공통 키워드 추출 방법은 기존의 방법에 비해 정확도가 약 3배가량 향상됨을 알 수 있고, 일반적인 웹 이미지 검색에서 가장 효과적으로 사용될 수 있는 캡션 후보 필터링 방법 역시 기존의 방법에 비해 정확도가 약 1.45배 향상되었음을 알 수 있다.



(a) 학습된 개념만을 포함하는 이미지에 대한 실험 결과



(b) 학습되지 않은 개념을 포함하는 이미지에 대한 실험 결과

<그림 3> 캡션 추출 결과의 정확도 비교

시스템에 학습되지 않은 개념을 포함하는 이미지에 대하여 실험을 하면 <그림 3>-(b)에서 보이는 것처럼 학습된 개념만을 포함할 경우 가장 좋은 성능을 보여줬던 공통 키워드 추출 방법이 매우 좋지 않은 성능을 보여주는 것을 확인할 수 있다. 따라서 학습되지 않은 개념이 존재할 수 있는 일반적인 WWW 환경에 적용하기 위해서는 캡션 후보 키워드 필터링 방법을 사용하는 것이 효과적임을 알 수 있다. 하지만 학습된 개념이 수가 늘어난다면 공통 키워드 추출 방법을 사용하는 것이 효과적일 것이다.

5. 결론

컴퓨터 기술 및 디지털 미디어 장치 등의 발달로 이미지와 같은 멀티미디어 데이터가 대량으로 생산되고 있으며 WWW 등을 통해 많은 사람들이 이를 서비스하거나 이용할 수 있게 됨에 따라 WWW에 존재하는 이미지 데이터에 대한 검색과 관리의 중요성이 점점 커지고 있다. 웹 이미지의 효과적인 검색을 위해서는 웹 이미지가 표현하는 개념을 잘 표현할 수 있는 캡션을 선택하여 이미지의 주석으로 사용해야 한다. 본 논문에서는 웹 페이지 상의 텍스트 정보만을 이용해서 캡션을 결정할 때 선택되는 캡션의 정확도가 낮아지는 문제점을 해결하기 위해서, 기존의 웹 이미지 마이닝 방법에 자동 이미지 주석 방법을 결합하여 최종 캡션을 선택하는 알고리즘으로서 '키워드 병합', '공통 키워드 추출', '개념 부류 키워드 필터링', 그리고 '캡션 후보 키워드 필터링' 을 제안하고 웹 이미지 캡션 추출 시스템을 구현하였다. 실험 결과에 의하면 일반적인 웹 이미지 검색 시스템에 캡션 후보 필터링 방법을 적용할 경우 기존의 방법에 비해 우수한 성능을 보여줄 수 있을 것이다.

참고문헌

[1] S.Rui, W.Jin, and T.Shua, "A Novel Approach to Auto Image Annotation Based on Pair-wise Constrained Clustering and Semi-nave Bayesian Model," *Proceedings of IEEE 11th International Conference on Multimedia Modeling*, pp. 322-327, 2005.

[2] J.Z.Wang and J.Li, "Learning-based Linguistic Indexing of Pictures with 2-D MHMMs," *Proceedings of the 10th ACM International Conference on Multimedia*, pp. 436~445, 2002

[3] Y.Mori, H.Takahashi, and R.Oka, "Image-To-Word Transformation based on Dividing and Vector Quantizing Images with Words," *Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[4] C.Cusano, G.Ciiocca, and R.Scettini, "Image Annotation using SVM," *Proceedings of Internet Imaging IV*, pp. 330~338, 2004.

[5] C.Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, pp.265~283, 1998.

[6] S. Sclaroff, L. Taycher, and M. La Cascia, "ImageRover: A Content-Based Image Browser for the World Wide Web," *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 2~9, 1997.

[7] J.R.Smith and S.F.Chang, "WebSeek: An Image and Video Search Engine for the World Wide Web," in *IS&T/SPIE Proceedings of Storage and Retrieval for Image and Video Database V*, pp. 84~95, 1997.

[8] C.Frankel, M.J.Swain, and V.Athitsos, *WebSeer: An Image Search Engine for the World Wide Web*, *Technical Report 96-14*, University of Chicago Computer Science Department, 1996.

[9] S.Mukerjea and J.Cho, "Automatically Determining Semantics for World Wide Web Multimedia Information Retrieval," *Journal of Visual Languages and Computing Vol.10*, pp. 585~606, 1999.

[10] N.C.Rowe and B.Frew, "Automatic Caption Localization for Photographs on World Wide Web Pages," *Information Processing and Management, Vol.34, No.1*, pp. 95~107, 1998.

[11] R.B.Yates and B.R.Neto, *Modern Information Retrieval*, Addison Wesley, pp. 74-84, 1999.