토픽 모델링을 이용한 이미지의 효율적인 표현방법

이바도O, 장병탁
서울대학교 컴퓨터공학부
e-mail : bdlee@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Efficient Method for Image Representation Using Topic Modeling

Bado LeeO, Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University

요 약

시각 피처를 사용한 이미지 표현은 이미지 검색 분야에서 이미 광범위하게 사용되고 있다. 특히 이미지 자체에 태깅이 되어있지 않거나 다른 추가 정보가 없는 경우에는 이미지 콘텐츠 자체의 정보만으로 검색하기 위해서는 이러한 전처리가 필수적이다. 이미지로 부터 얻어진 시각적 피처들이 시각 단어로 사용되기 위해서는 *k*-means 와 같은 군집 알고리즘을 통한 시각적 피처의 양자화를 위한 전처리가 필요한데, 시각 단어의 개수 *k*를 정하는데 모호함이 있다. 본 논문에서는 임의의 *k*를 사용하더라도, 대표적 토픽 모델링 기법인 LDA (Latent Dirichlet Allocation) 를 사용하여 데이터의 차원을 줄이게 되면 여러개의 시각적 단어들의 조합을 각각의 토픽이 나타낼 수 있게 됨을 이미지 검색 성능으로써 확인해 보고, 이러한 방법을 사용하면 표현형의 사이즈를 줄일 수 있고, 검색에 있어서도 이미지의 유사성을 더욱 효과적으로 표현할 수 있음을 확인해 본다.

1. 서 론

정보의 범람 시대에 살면서 원하는 데이터를 빠른 시간 내에 찾는 것이 점점 더 어려워 지고 있다. 인터넷에는 하루가 멀게 대용량의 이미지가 업데이트 되고 있기 때문에 이미지의 경우 이 문제는 더 심각하다. 이미지 검색의 경우 대부분 태그 정보를 사용하고 있다. 태그 정보는 이미지 자체에 비해서는 그 용량이 작기 때문에 빠른 검색을 할 수 있지만, 태그 정보가 잘못 붙어 있거나, 새로운 이미지의 경우에는 사람이 바로 작업을 하지 않는 경우에는 아예 태그 자체가 없는 경우도 많이 있다.

태그 없이 이미지 검색을 하는 방법에는 이미지 콘텐츠 기반 검색이 있다. 일반적으로 알려진 가장 간단한 방법은 다음과 같다. 우선 SIFT와 같은 방법을 사용하여 이미지에서 피처 값들을 뽑아낸다. 그리고 *k*-means와 같은 군집 알고리즘을 사용하여 이미지를 시각 단어의 히스토그램으로 표현한다. 이 방법의 경우 적절한 *k*의 값을 정하는데 어려움이 있다. 왜냐하면 데이터에 따라 적절한 *k*의 값이 서로 다르기 때문이다.

이 논문에서는 이미지 표현형의 높은 표현력을 위해 *k*의 값을 크게 잡은 후에 LDA[1-3] 를 사용하여 차원을 줄여 이미지를 검색하는 방법을 제안한다. 이러한 방법을 사용하면 작은 크기의 *k* 값을 사용하는 것보다 비교적 나은 성능을 얻을 수 있음을 확인할 수 있었다. LDA를 사용하면 적은 수의 차원을 가지고도 이미지를 효과적으로 표현할 수 있다.

2. 관련연구

2.1 토픽 모델 기반의 이미지 표현

근래에 확률적 문서 모델링 기법이 텍스트 검색 분야에서 성공적으로 적용되고 있다[4]. 이러한 모델에서 이미지는 토픽의 혼합 모델로 표현된다. 이러한 토픽들은 한 문서 내에서의 단어들을 연계시키고 또한 서로 다른 문서들 간의 연계성도 모델링 할 수 있다. 이러한 접근 방법은 이미지 표현에서도 널리 사용되고 있는데, 이미지 분류[5]나 비전에서의 물체 분류[6] 분야에서 성공적으로 활용되고 있다.

2.2 LDA 를 사용한 이미지 검색

LDA 를 이용한 이미지 검색은 이미 많은 연구가 진행되고 있다. 텍스트 기반의 이미지 검색으로는 이미지에 태깅 되어있는 텍스트 데이터를 LDA 로 처리하여 검색에 사용하거나[7] 이미지 콘텐츠 기반의 검색으로는 대

용량 데이터베이스로부터의 검색[8] 뿐만 아니라, 지속적으로 증가하는 방대한 이미지 데이터에 대한 자동 태깅[9] 도 연구가 진행되고 있다.

몇몇 연구에서는 상당히 큰 데이터베이스에서의 검색[8]에 이러한 방법이 응용되고 있는데, 토픽 모델링을 사용하면 이미지 뭉치의 표현 사이즈를 효과적으로 줄일 수 있다. 하지만 단순히 콤팩트한 이미지의 표현만을 위해서라면 이러한 방법이 필수적이지는 않다. 왜냐하면 시각적 피처들을 군집 알고리즘으로 양자화하는 과정에서 군집의 갯수를 작게 설정함으로써, 이미지의 콤팩트한 표현형을 얻을 수 있기 때문이다.

앞에서 설명한 연구들에서는 *k*-means 의 *k* 값을 작게 설정하여 이미지를 표현한 경우와 *k*값을 크게 잡은 후 LDA를 사용하여 같은 차원의 표현형으로 이미지를 변환했을 때의 경우를 비교하지 않았다. 이 논문에서는 *k*-means 만 사용하여 이미지 표현한 경우와, *k*-means 후 LDA 를 사용한 경우 이미지 검색에서 어떤 효과가 있는지를 비교 실험을 통하여 확인해 보고자 한다.

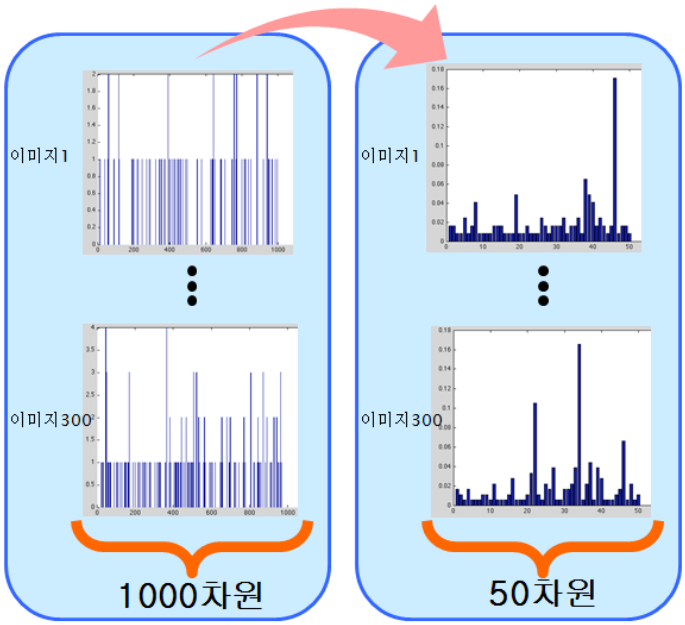


그림 1 LDA 처리 전과 후의 이미지 표현형 비교

3. 토픽 모델 기반 이미지 표현형 생성과 유사도 측정

일반적인 이미지의 벡터형 변환은 여러가지 방법이 있지만, 이미지로 부터 뽑아진 시각적 피처들을 군집 알고리즘을 사용하여 시각적 단어들로 변환한 후에 이를 이용하여 이미지를 다시 시각적 단어의 히스토그램으로 나타내는 방법이 널리 사용된다. 이런 방법에서는 시각적 단어의 개수를 정하는 것이 중요한 문제가 될 수 있는데, 시각적 단어의 개수에 따라 표현형의 표현력이 달라질 수 있기 때문이다. 따라서 이 논문에서는 큰 수의 시각적 단어를 우선 이용하여 일단 이미지를 히스토그램으로 변환한 후에, 이를 bag-of-word로 LDA 의 입력 값으로 사용하여 최종적으로 각 토픽의 분포 확률로써 각각의 이미지 표현형을 얻는 방법을 제안한다.

각 이미지 간의 유사도 측정은 JS(Jensen-Shannon) divergence를 사용하였다. JS divergence *JS(p,q)* 는 다음과 같은 식으로 나타난다.

$$JS(p,q) = \frac{1}{2}\left(KL\left(p,\frac{p+q}{2}\right) + KL\left(q,\frac{p+q}{2}\right)\right)$$

$$KL(p,q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i}$$

JS divergence는 두 개의 확률 분포 간의 유사도를 측정하는 함수로써, 통계학과 정보이론 분야에서 널리 사용되며, 대칭적이지 않은 KL(Kullback-Leibler) divergence를 기반으로 하여 대칭적이고 유한한 값을 출력하도록 한다.

4. 실험결과 및 분석

4.1 데이터 및 전처리

이 컨텐트 베이스 실험에서의 목적은 쿼리로 주어진 이미지와 최대한 비슷한 이미지들을 검색하는 데 있다. 실험은 총 240 개의 이미지에 대하여 수행되었다. 이미지들은 총 8 개의 카테고리에 해당하게 되는데 이중 7 개의 카테고리의 이미지는 Caltech-101 의 데이터 이미지를 사용하였다. 이들 데이터는 특정 기준을 가지고 선택되었다. 이 논문에서의 실험이 단순히 이미지 검색 성능을 측정하기 위함이 아니라, 각 표현형들이 검색 성능에 어느정도 영향을 미치는지 알기 위하여 수행되었기 때문에 이미지들은 최대한 노이즈가 없는 (예를 들어 배경이 단순하고 주 사물이

표 1 각 카테고리 쿼리에 대한 이미지 검색 실험 결과

카테고리	Accordion	Dollar bill	Face	Garfield	Metronome	Starfish	Stop sign	Sunflower	평균값 (표준편차)
k-means (*k*=1000)	24%	25%	91%	34%	28%	31%	45%	12%	36% (24)
k-means (*k*=50)	51%	59%	99%	54%	45%	48%	54%	64%	59% (17)
LDA Applied to *k*-means (T=50,*k*=1000)	69%	78%	94%	61%	49%	54%	72%	68%	68% (14)

비슷하게 생긴) 것들을 골라서 실험에 사용하였다. 나머지 1 개의 카테고리는 같은 배경에서 촬영한 저자의 사진을 사용하였다. 8 개의 카테고리는 각각 다음과 같다: 아코디언, 달러 지폐, 얼굴 사진, 가필드, 매트로놈, 불가사리, 정지 표지판, 해바라기.

k-means 중심점의 라벨을 시각적 단어로 사용하는 방식은 컨텐트 베이스의 이미지 검색 분야에서 널리 사용된다. 이러한 접근을 위해서 우선 이미지에서 시각적 피처들을 추출해낼 필요가 있다.

이 논문에서는 잘 알려져 있고 널리 사용되는 SIFT 를 사용하여 이미지들로부터 시각적 피처들을 추출하였다. SIFT[10] 디텍터는 모서리와 같은 특이점들을 이미지의 가우시안 피라미드의 층간 차이값을 이용하여 찾아낸다. 이렇게 찾아진 특이점들은 디스크립터에 의해 128 차원의 gradient-based 벡터 값으로 전환되게 된다. 이 논문에서 사용한 이미지들에서는 장당 200 에서 300 개의 시각적 피처들이 추출되었다. 정확도는 전체 학습 데이터 셋에 대한 검색을 수행하여 측정하였다. 각각의 카테고리에 대하여 이러한 과정이 수행되었고 전체 값의 평균이 정확도로써 사용되었다.

실험은 총 3가지로 진행하였다. 우선 *k*-means 로만 전처리를 하여 가장 높은 성능을 나타내는 *k*값을 찾았다. 이 *k* 값을 이용하여 이미지 검색 수행 성능을 측정하였다. 두번째 실험은 아주 높은 값의 *k* 값을 이용하여 동일한 실험을 수행하였다. 세 번째로는 이 높은 값의 *k* 값을 이용하여 전처리를 한 후에 다시 LDA 로 데이터를 처리한 경우의 이미지 검색 성능을 측정하여 그 값을 비교해 보았다. 이때 LDA 의 토픽의 수는 비교를 위하여 *k*-means 의 가장 좋은 결과의 *k*값과 똑같이 설정하였다.

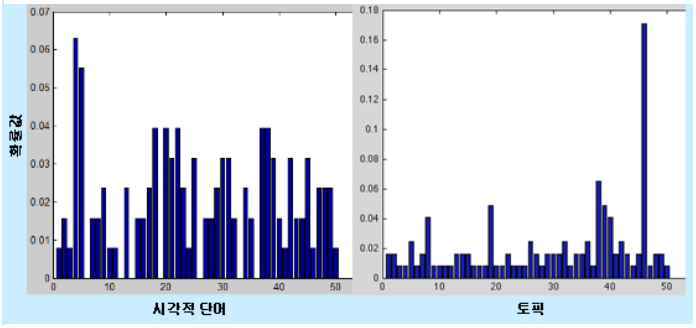
4.2 토픽 모델 적용 전후 검색 성능 비교

세 가지 실험의 결과는 표 1 에 나타나 있다. *k*=1000 인 *k*-means 만을 사용하여 검색을 수행하였을 때 평균적인 정확도는 36%이다. 이 결과는 다른 두 결과에 비해서 매우 낮은 결과이다. 이는 최적의 *k* 값을 찾지 않고 곧바로 큰 값의 *k* 를 곧장 사용하기 때문인데, 최적의 *k* 값인 *k*=50 일 때는 이미지 검색 정확도가 59% 로 월등히 나아진다. 비교적 결과가 좋지 않은 *k*=1000 으로 얻어진 데이터를 LDA 로 처리할 경우는 결과가 68%로써 최적의 *k* 를 사용한 경우보다 더 좋은 성능을 보여주고 있다.

4.3 결과 분석 및 논의

그림 1 왼쪽에 있는 이미지의 시각적 단어 히스토그램을 관찰해 보았을 때, 벡터가 매우 성긴 것을 알 수 있다. 한정된 개수의 단어가 큰 차원에 할당됨으로써 같은 단어로 분류되었어야 할 시각적 피처들이 두개의 단어로 분리되는 결과가 생긴다. 결과적으로 너무 큰 차원을 사용하게 되면 이미지들 간의 유사성을 판단하기 어렵게 된다.

그림 2 토픽 모델을 적용한 이미지 검색 예



(a) *k*-means (b) *k*-means LDA

그림 3 *k*-means 와 LDA 의 50차원에서의 표현형 비교

작은 값의 *k* 를 사용하게 되면 앞서 말한 시각적 피처가 두개의 단어로 분리되는 현상이 발생하지 않는다. 하지만 여기서도 다른 문제점이 발생하는데, 그것은 서로 다른 두개의 시각적 단어로 구분되어야 할 시각적 피처들이 하나의 피처로 판별된다는 것이다. 다르게 말하면 모델의 표현력이 데이터에 비해 부족하다고 말할 수 있다. 만약 데이터의 양이 늘어난다면 이러한 문제는 더욱 부각될 것 이고 성능은 데이터의 증가와 더불어서 급격하게 감소될 것이다.

이러한 문제들을 극복하기 위해서 데이터의 사이즈에 따라 다른 크기의 *k* 값을 사용하는 것이 한 방안이 될 수 있는데 이러한 방법을 사용할 경우 적절한 *k* 의 크기를 어떻게 찾을것인가 하는 또 다른 문제가 생길 수 있다. 또한 다른 크기의 데이터 셋을 표현할 때마다 다른 크기의 *k* 값을 사용하게 되면 서로 다른 데이터 셋의 해당 이미지들을 비교하기 위하여 또 다른 방법을 찾아야 한다는 문제가 생긴다. 뿐만 아니라 적절한 가장 높은 성능의 *k* 값을 사용한다 하더라도 여전히 서로 다른 시각적 피처들이 하나로 뭉치는 위험을 배제할 수 없다.

마지막 실험의 경우 그 성능이 68%로 *k*-means 를 이용한 최적의 결과보다 더 좋은 성능을 보이고 있는데, 이러한 결과를 보이는 이유는 LDA 의 토픽들을 시각적 단어로 간주할 경우 각 토픽들이 문제공간의 기저 벡터와 같이 작용하기 때문이다. 아주 큰 *k* 값을 사용하여 시각적 피처들을 서로 다른 것으로

구분하였다고 하더라도, LDA 의 토픽들이 다시 이러한 조각들을 합치는 역할을 하는 것이다.

그림 2 는 이미지 검색 결과 예이다. 실험 결과를 보면 달러 지폐를 제외하고는 좋은 성능을 보이고 있다. 달러 지폐 이미지의 경우 대부분의 이미지가 노이즈가 많이 섞여 있기 때문이 좋은 결과를 얻기 힘들었다. 이 논문의 실험에서는 시각적 피처를 뽑기 전에는 다른 어떠한 전처리도 수행하지 않았기 때문에 배경 제거 같은 비전 전처리를 통하면 이 실험보다 나은 결과를 기대할 수 있다. 또한 앞에서 언급하였듯이 실험을 위하여 선택된 이미지 데이터 셋은 달러 지폐 셋을 제외하고는 서로 비슷비슷한 이미지를 사용하였다. 자연스런 데이터 셋을 구분하려면 좀더 새로운 방법론이 필요할 것이다. k-means 와 k-means - LDA 각각에 의해 변환된 히스토그램을 비교해 보면 성능 차이의 이유를 좀더 확실히 이해할 수 있다. 그림 3 은 각각의 표현형의 대표적인 예이다. 왼쪽이 k=50 인 k-means 로만 처리된 데이터의 히스토그램이고 오른쪽이 k=1000 인 k-means 처리 후 LDA 를 거친 히스토그램이다. 왼쪽의 그림은 값의 변화 폭이 작은데, 이는 표현력을 현격하게 저해하는 요소로 작용한다. 반면에 k-means-LDA 로 처리된 이미지 히스토그램의 경우 그 높이의 변화폭이 크고 표현할 수 있는 문제공간의 넓이도 k-means 만 사용한 경우보다 클 것임을 예측할 수 있다.

5. 결론

본 논문에서는 이미지를 시각적 단어로 표현하는 방법에 있어서 LDA 를 사용 할 경우 어떤 효과가 있는지 실험적으로 살펴보았다. k-means 만 가지고 이미지 데이터를 처리 하였을 때, k 의 크기가 이미지 검색 성능에 큰 영향을 미치게 된다. 곧바로 적은 수의 시각단어들 만으로 이미지를 처리하는 경우에는 서로 다른 시각적 단어가 같은 것으로 처리되기 때문에 이미지에 대한 표현 능력이 떨어지게 된다. 반대로 곧장 큰 값의 k 를 사용하면, 동일하게 간주 되어야 할 시각적 단어가 분리되어 버리는 현상이 발생해서 이미지간의 유사도를 측정하는 데 어려움이 있다. 또한 최적의 k 값을 찾는 것도 시간이 많이 걸리게 된다. 또한 최적의 변수를 찾는다 하더라도, 여전히 서로 다른 시각적 피처들이 뭉쳐 버리는 결과는 피할 수 없다. 하지만, 임의의 큰 k 값을 사용한 후에 이를 LDA 로 처리해 주게 되면 토픽 모델의 각각의 토픽들이 시각 단어들을 추상화 하여 그 결합 관계를 내포할 수 있기 때문에 전처리 단에서의 k 값에 상관없이 안정적인 이미지 검색 결과를 얻을 수 있게 된다.

감사의 글

이 논문은 교육과학기술부의 재원으로 국가연구재단의 지원을 받아 수행된 연구(0421-20110032, 지능형 추천 서비스를 위한 인지기반 기계학습 및 추론기술, Videome)이며, 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발) 및 교육과학기술부의 BK21-IT사업에 의해 일부 지원되었음.

참고문헌

[1] T. Griffiths, and M. Steyvers, Finding scientific topics, *PNAS*, vol.101, pp.5228-5235, 2004.

[2] A. Vedaldi, and B. Fulkerson, VLFeat : An open and portabl library of computer vision algorithm, *http://www.vlfeat.org/*, 2008.

[3] D. Blei, A. Ng, and J. Michael, Latent Dirichlet allocation, *JMLR*, vol.3, pp.993-1022, 2003.

[4] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, Probabilistic auther-topic models for information discovery, *SIGKDD*, pp.306-315, 2004.

[5] A. Bosch, A. Zisserman, and X. Munoz, Scene classification via pLSA, *ECCV*, vol.3954, pp.517-530, 2006.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, Learning object categories from google's image search, *ICCV*, pp.1816-1823, 2005.

[7] D. Blei, M, David, and M. Jordan, Modeling annotate d data, *SIGIR*, pp.127-134, 2003.

[8] E. Horster, R. Lienhart, and M. Slaney, Image retrieval on large-scale image database, *CIVR*, 2007.

[9] J. Jeon, V. Lavrenko, and R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, *SIGIR* , pp.119-126, 2003.

[10] D.G Lowe, Object recognition from scale-invariant features, *ICCV*, vol.2, pp.1150, 1999.