내용 정보를 이용한 이미지 자동 태깅

장현웅, 조수선
한국교통대학교
e-mail : jhwskorg@gmail.com

Automatic Annotation of Image using its Content

Hyun-Woong Jang, Soosun Cho
Dept. of Computer Science & Information Engineering, Korea National University of Transportation

요 약

이미지 인식과 내용분석은 이미지 검색과 멀티미디어 데이터 활용 분야에서 핵심기술이라 할 수 있다. 특히 최근 스마트폰, 디지털 카메라, 블랙박스 등에서 수집되는 영상 데이터 양이 급격히 증가하고 있다. 이에 따라 이미지를 인식하고 내용을 분석하여 활용할 수 있는 기술에 대한 요구가 점차 증대되고 있다. 본 논문에서는 이미지 내용정보를 이용하여 자동으로 이미지로부터 태그정보를 추출하는 방법을 제안한다. 이 방법은 기계학습 기법인 CNN(Convolutional Neural Network)에 ImageNet 의 이미지 데이터와 라벨을 학습시킨 후, 새로운 이미지로부터 라벨정보를 추출하는 것이다. 추출된 라벨을 태그로 간주하고 검색에 활용한다면 기존 검색시스템의 정확도를 향상시킬 수 있다는 것을 실험을 통하여 확인하였다.

1. 서론

최근 스마트폰, 디지털 카메라, 블랙박스 등의 발전으로 영상 데이터의 수집 양이 급격히 증가하고 있다. 이에 따라 플리커, 페이스북, 인스타그램과 같은 대용량 소셜 미디어 공유 사이트가 급속하게 발전되었고 사용자들은 언제 어디서든 인터넷에 접속해서 영상데이터를 공유할 수 있게 되었다. 많은 양의 이미지 데이터가 웹 공간에 저장됨에 따라 정확한 인덱싱과 이미지 탐색 기법이 중요하게 되었다.

다양한 이미지 탐색 기법 또한 활발하게 연구되고 있는데 그 중 가장 기본적인 접근 방법은 키워드 기반의 검색이다. 키워드 검색은 사용자가 이미지에 의미 있는 태그를 붙여 이미지 검색에 활용하는 것이다. 하지만 폭소노미 기반의 웹 이미지에는 그 이미지의 내용과 관련 없거나 사용자의 주관적인 판단으로 추가되는 태그들이 포함되기 때문에 정확한 이미지 검색에 한계가 있다. 이를 극복하기 위하여 태그들의 의미적 중요도를 분석하고 이미지 검색에 활용하려는 연구[1]가 있었다.

이미지 내용기반 검색 기법은 키워드 또는 태그기반의 검색과는 다른 차원의 이미지 탐색 기법이다. 인터넷 및 컴퓨팅 속도가 급속히 빨라짐에 따라 이미지 내용기반 검색에 대한 요구가 증대되고 있고 다양한 연구가 진행되고 있다[2]. 최근에는 이미지 인식과 추출 기술이 뛰어난 성능을 보이면서 이미지 내용 기반의 자동 태깅 기법이 주목 받고 있다[3].

이미지 태깅은 웹이나 스마트 기기에서 이미지를 공유할 때 검색을 위해 사용되는 기법이라고 볼 수 있다. 수동적인 이미지 태깅은 사용자의 주관적인 요소가 포함이 될 수 있고, 특히 모바일, 디지털 카메라와 같은 기기에서는 사용자가 직접 태그를 달아야 하는 많은 불편함이 따른다. 본 연구에서는 이미지 내용기반의 인식기법을 사용해서 이미지에 자동으로 태그를 붙여 이미지 검색의 정확도를 높이고, 직접 태깅을 하는 시간과 비용을 줄일 수 있는 방법을 제안한다. 실험 데이터는 이미지를 기반으로 하는 SNS 의 일종인 인스타그램을 사용한다. 사진, 영상 중심의 소셜미디어인 인스타그램은 페이스북이 인수하면서 2014 년 실사용자 수가 트위터를 사용자 수를 넘어 인스타그램을 활용한 다양한 연구가 진행되고 있다[4]. 특히 사용자들이 직접 올리는 이미지로서 정형화되어 있지 않고 태그가 정확하지 않은 이미지들이 많이 있어 본 연구에서 정확도를 증가 시킬 수 있는 대상으로 사용되었다.

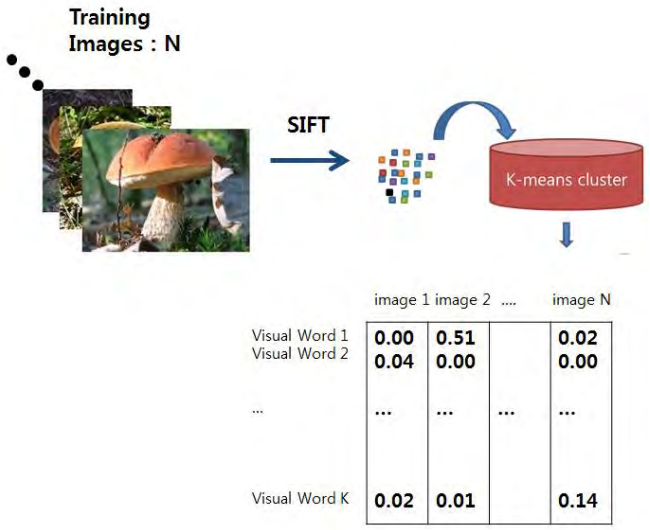
본 논문의 구성은 다음과 같다. 2 절에서는 관련된 연구에 대해 소개하고 3 철에서는 본 논문에서 제안하는 방법을 사용한 실험을 소개한다. 4 절에서는 결과를 분석하여 평가하고 5 절에서는 결론을 맺는다.

2. 관련연구

2.1 시맨틱 주석의 자동 생성

이미지 내용기반의 분류에서는 BoVW(Bag of Visual Words)방법이 다양한 곳에서 사용되고 있다. BoVW 기법에서는 이미지에서 추출된 특징점들을 K-means 로 군집화하여 Visual Words 를 생성한다. 각각의 이미지를 Visual Word 벡터로 표현하여 SVM(Support Vector Machine)기반의 분류기에 학습을 시키고 분류하는 것이 기본적인 과정이다. 특히 어떤 특징을 추출하느냐

에 따라서 분류의 목표나 성능이 달라진다. SIFT 특징점 추출 알고리즘은 색상, 위치, 크기, 회전 등의 변화에 강인한 이미지 특징점 추출 알고리즘이다. SIFT 알고리즘과 BoVW 기법을 사용해서 이미지에서 내용정보를 추출해 이미지에 주석을 달아주는 연구가 있었다[5].



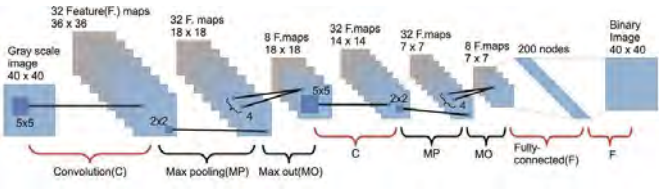
(그림 1) BoVW 기법에서의 이미지 표현 방법

SIFT 알고리즘과 K-means 를 통해 얻은 Visual Words 의 집합을 local visual blocks 라고 할 수 있다. 그림 1 의 과정을 거쳐 추출된 local visual blocks 특징 R_i 와 텍스트 집합으로 구성된 주석 keyword K_j 사이의 조건부 확률 $P(K_j|R_i)$을 통해 적합한 주석을 얻을 수 있다. 관련연구[5]에 따르면 이 방법을 통해 이미지 검색에서 정확도를 높일 수 있었다. 한편, SIFT 는 Object 분류에서 강점을 보이지만 Scene 과 같은 배경 이미지 분류에서 정확도가 떨어지는 단점이 있다[2]. 최근에는 SIFT 알고리즘으로 특징을 추출하여 분류하는 방법보다 CNN(Convolutional Neural Network)과 같이 깊은 신경망 네트워크(deep learning)를 사용하여 이미지를 분류하는 방법이 더 좋은 성능을 보이고 있다[6].

2.2 깊은 신경망 네트워크를 사용한 내용 분석

최근에 영상 인식 대회[7]에서 CNN(Convolutional Neural Network)은 다른 알고리즘보다 물체 분류 및 인식에 좋은 성능을 보이고 있다[8]. CNN 은 다양한 물체 인식 분야에 적용되어 왔다. 이미지 인식뿐만 아니라 음성인식, 영상인식에서도 뛰어난 성능을 보이고 있다. 토론토 대학에서 대규모 이미지 검색을 위해 CNN 기술을 개발하여 이미지 인식에서 뛰어난 성능을 보였다[8]. 이 시스템은 1,000 개의 카테고리 클래스로 구성되어 있으며 120 만개의 이미지로 훈련시켜 Net 파일(Network)을 만들었다. 이는 영상인식에서 약 85%의 정확도를 보였다[8]. 하드웨어가 발전하고 데이터 양이 증가함에 따라 CNN 을 사용한 다양한 분야에서 기존의 최고 성능을 능가하는 결과를 보이고 있다. 특히 인식분야에서 뛰어나다. 예를 들어, 손 글씨 인식, 표지판 인식, 교통신호 인식, Caltech101 이미지 인식, ImageNet 이미지 인식 등 Caltech101 이미지 인식을 제외한 모든 곳에서 기존기록보다 뛰어난 기록을 세우고 있다[9]. 이미지가 적은 데이터 집합을 사용할 때보다 이미지 데이터가 많은 집합을 사용할 수록 더 좋은 성능을 보이는 것을 알 수 있다[9].

깊은 신경망 네트워크의 중심인 CNN 은 사람이나 동물의 시각 처리 과정을 모방하기 위해 개발된 신경망이다. CNN 의 구조는 그림 2 와 같이 크게 세 가지 Convolution(C)계층, pooling(M)계층, Fully-connected(F)계층으로 이루어진다[10].



(그림 2) Convolutional Neural Network 구조

CNN 은 신경망 네트워크로 하위 계층부터 Convolution 으로 특징을 추출하고 pooling 으로 추상화하는 것을 반복하면서 점차 높은 수준의 특징을 추출하게 된다. CNN 의 최상위 계층은 Fully-connected 계층으로 구성되는데 이전 계층에 있는 모든 뉴런의 결과를 받아 최종 결과를 계산한다. 각 계층은 2 차원적으로 배열된 특징맵(feature map)으로 구성된다. 특징맵은 용도에 따라 1 차원, 2 차원, 3 차원 등으로 정할 수 있다[9].

Convolution 계층은 전 계층의 결과로부터 복수의 입력을 받아 같은 노드에서 공유된 가중치 연산을 하는데, 가중치는 Convolution 마스크 연산 처리를 말한다[11]. 각 Convolution 계층을 수행한 후 다음으로 Pooling 계층을 수행한다. Pooling 계층은 Convolution 계층과 동일한 수의 노드를 가지며 1:1 로 연결된다. Pooling 으로 Max-Pooling 을 사용하는데 pooling 하는 블록 내의 최대값을 취하는 역할을 한다[11]. 이는 특징이 되는 값은 보존하고 크기를 줄여 연산이 빠르게 될 수 있도록 돕는 역할을 한다. Convolution 과 Max-Pooling 을 한 후에 최상단에 있는 Fully-Connected 를 수행하게 된다[10]. Fully-Connected 는 전 계층의 모든 뉴런들과 연결되어 최종 인식 결과를 결정한다. Fully-Connected 계층 후에는 Convolution 과 같은 연산 계층이 없다.

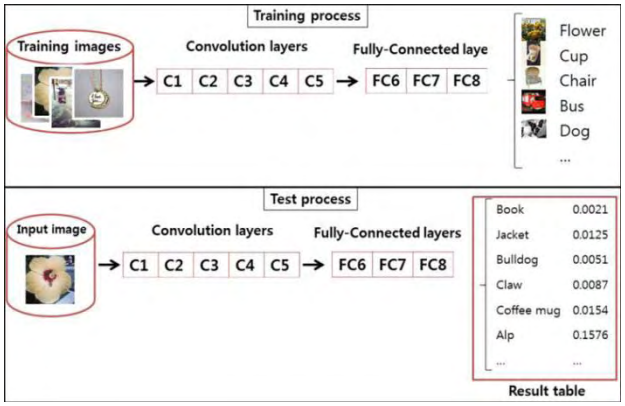
Convolution 계층에서 정확하고 의미 있는 특징 추출하는데 이것은 학습을 통해 이루어 진다. 학습은 오류 역전파 알고리즘(Backpropagation pass)과 경사도 연산 방법(Computing the Gradients)을 통해 이루어 지는데, 합성함수 미분법인 연쇄법칙(Chain Rule)을 사용하여 Convolution 계층과 Pooling 계층을 연결한다[8]. 두 계층 사이의 노드 간 연결은 제한된 연결 방법을 사용한다[8]. 이는 연산속도를 증가시키고 오차를 줄이기 위한 방법이다.

3. 구현 및 실험

구현에서는 토론토대학에서 구현한 Net 모델을 참조했다. 훈련된 Net 모델은 Krizhevsky[8]의 모델을 기초로 디자인되었다.

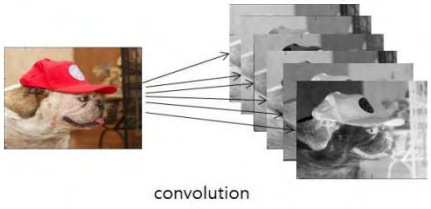
이미지의 테스트 데이터로 최근 사용자 수와 영상 데이터가 급증하고 있는 인스타그램의 이미지를 활용했다. 인스타그램은 일상생활 속에서 발생되는 사건을 스마트 기기를 사용해서 촬영하고 이를 자신의 인스타그램에 게시하는 미디어 공유 SNS 이다[4]. 인스타그램의 이미지들은 정제되지 않은 이미지들이고 실제 생활에서 많이 사용되는 이미지이기 때문에 정확한 검색이 쉽지 않다. 본 실험으로 이미지 검색의 정확도를 증가시킬 수 있다는 것을 증명하기 위해 인스타그램의 이미지 데이터를 활용하여 15 개의 카테고리 별로 30 개씩의 이미지를 검색했다. 즉, 총 450 개의 이미지로 테스트를 수행하였다.

이미지 훈련 데이터로 1,000 개의 클래스 카테고리와 120 만개로 구성된 ImageNet2012 를 사용해서 훈련된 신경망(Net)을 구성했다[7]. 전체적으로 5 개의 convolution 계층과 3 개의 Fully-Connected 계층으로 구성했다. convolution 계층에 이미지를 feeding 하기 전에 각 이미지를 256×256 으로 크기로 조정한 후에 이미지에서 224×224 픽셀을 랜덤으로 추출했다. RGB 값의 이미지 픽셀을 신경망에 훈련시킬 수 있었다.



(그림 3) 전체적인 Network 구성 과정

먼저 Convolution 을 하는 이유는 이미지의 전체적인 정보를 모두 고려하기 위해서이다. 그렇기 때문에 SIFT 나 SURF 보다 배경이미지에 강한 성능을 보이고 있다. Convolution 을 여러 번 하는 이유는 convolution 을 하면서 이미지의 정보들이 겹치게 되는데, 이때 이미지의 특징을 불변하게 학습할 수 있기 때문이다. 인식하고자 하는 물체가 항상 가운데 위치하고 있지 않을 가능성을 고려한 것이다.



(그림 4) convolution 의 예

N×N 의 입력 이미지가 Convolution 계층에 연결되어 있고, m×m 의 Convolution 필터가 있을 때 Convolution layer 의 크기는 (N-m+1)×(N-m+1)이 된다. 각 Convolution 계층은 이전 계층의 출력으로부터 x_{ij}^{l} 를 입력을 받는다.

$$for\ i,j = 0,1,\dots,N-m$$

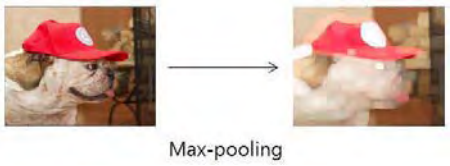
$$x_{ij}^{l} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} w_{ab}\, y_{(i+a)(j+b)}^{l-1} \quad (1)$$

$$y_{ij}^{l} = f\left(x_{ij}^{l}\right),\ \ f\text{는 nonlinear function} \quad (2)$$

식 (1)은 이전 계층의 convolution 필터 가중치(w_{ab})를 계산한 출력으로부터 입력을 받아 합한 값이다. 식 (2)는 이전 계층에서 출력된 값을 활성화 f 함수로 계산한 값이다. 여기서 f 함수는 활성화 함수인 nonlinear 함수를 말하는데 식(3)과 같은 ReLU(Rectified Linear Unit)함수를 사용했다[8].

$$f(x) = \mathrm{f}\max(0,\mathrm{x}),\ f(x) = \begin{cases} x & if\ x > 0 \\ 0 & otherwise \end{cases} \quad (3)$$

Convolution 계층 다음으로 1:1 로 연결된 Max-pooling 계층을 지나게 되는데 Max-pooling 은 특별한 연산 없이 최대 평균값을 추출해 연산을 빠르게 하고 이미지의 노이즈를 줄이고 공간적인 불변성을 높여 정확도를 높인다.



(그림 5) Max-Pooling 의 예

마지막으로 Fully-Connected 계층을 수행하게 되는데, FC6 와 FC7 계층은 dropout 을 사용해서 정규화를 시키고 마지막 계층에서는 soft-max 분류를 수행한다[10].

4. 결과분석 및 평가

<표 1> 일반 사진에 붙은 잘못된 태그와 CNN 을 통해 태그를 붙인 결과 비교

Image to be annotated	Instagram Manual tag	Automatic tags using CNN
	bike	Chain, necklace
	bike	Motor scooter, scooter

	house	Plate, cheeseburger

<표 1>에서 잘못된 태그를 가진 사진의 예들을 볼 수 있다. CNN 을 통해 사진을 분석하고 태그를 붙인 결과, 기존에 붙은 태그보다 더 정확한 태그가 붙여진 것을 알 수 있다. 사진과 CNN 으로 자동 태그된 결과를 보면 태그명이 정확하고 조금 더 자세한 설명이 추가되어 있는 것을 볼 수 있는데 이것은 CNN 을 사용함으로써 검색의 정확도와 함께 섬세함을 향상시킬 수 있다는 것을 알려주고 있다.

<표 2> 각 카테고리 별로 검색된 인스타그램 태그와 CNN 으로 재 태깅한 태그 결과

Categories	instagram		CNN	
	correct tags	Rates (%)	correct tags	Rates (%)
bike	20	66.67	23	76.67
bird	15	50.00	21	70.00
bottle	23	76.67	21	70.00
building	23	76.67	15	50.00
bus	8	26.67	21	70.00
car	23	76.67	28	93.33
chair	19	63.33	18	60.00
cup	19	63.33	27	90.00
flower	18	60.00	18	60.00
food	12	40.00	22	73.33
house	7	23.33	17	56.67
mountain	17	56.67	21	70.00
pet	24	80.00	22	73.33
roadsign	27	90.00	22	73.33
sea	19	63.33	21	70.00
합계	274	60.89	317	70.44

<표 2>는 인스타그램에서 각 카테고리 별로 30 개의 이미지를 검색해서 태그와의 적합성을 검사한 결과와 CNN 을 통해서 다시 태그를 붙여 적합성을 검사한 결과를 비교한 것이다. 총 450 개의 이미지를 가지고 적합성 검사를 했다. 일반적으로 사용자가 붙인 태그인 인스타그램 검색결과는 전체적으로 60.89%의 태그 정확도가 나왔다. 이를 CNN 으로 다시 태그를 붙인 결과 사진과 태그의 적합도가 70.44%로 약 10% 가량 증가한 것을 볼 수 있다.

5. 결론

본 연구에서는 CNN 을 활용한 이미지 자동 태깅 방법을 제안하고, 최근 사용자와 영상 데이터가 급증하고 있는 소셜 미디어 공유 사이트인 인스타그램의 이미지들로 실험했다. 이는 정형화된 이미지를 사용하지 않고 사용자들이 실생활에서 올리는 이미지를 사용함으로써 실제 대용량 이미지 공유 사이트에 활용이 될 수 있음을 보이기 위한 것이다.

이번 실험에서는 CNN 을 통해 인스타그램 태그 정확도를 약 10% 향상시킴으로써 검색의 정확도와 섬세함을 증가시킬 수 있다는 것을 증명했다. CNN 관련 기술은 지속적으로 발전하고 있으므로 향후 이를 활용한 다양한 연구를 진행할 계획이다.

참고문헌

[1] S. J. Lee and S. Cho, “Tagged Web Image Retrieval Re-ranking with Wikipedia-based Semantic Relatedness”, Journal of Korea Multimedia Society, Vol.14, No.11, pp.1491-1499, 2011

[2] H. W. Jang and S. Cho, “Image Classification Using Bag of Visual Words and Visual Saliency Model”, KIPS Transactions on Software and Data Engineering, Vol.3, No.12 pp.547-552, 2014

[3] J. W. Ha, B. H. Kim, B. Lee and B. T. Zhang, “Auto-tagging Method for Unlabeled Item Images with Hypernetworks for Article-related Item Recommender Systems”, Journal of the Korean Institute of Information Scientists and Engineers: Computing Practices and Letters, Vol.16, No.10, pp.1010-1014, 2010

[4] M. J. Nam, J. I. Kim and J.H. Shin, “A User Emotion Information Measurement Using Image and Text on Instagram-Based”, Journal of Korea Multimedia Society, Vol. 17, No. 9, pp.1125-1133, 2014

[5] M. Liang, J. Du, V. Jia and Z. Sun, “Image Semantic Description and Automatic Semantic Annotation”, Control Automation and Systems (ICCAS), 2010 International Conference on, pp.1192-1195, 2010

[6] P. Fischer, A. Dosovitskiy and T. Brox, “Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT.”, CoRR, abs:1405.5769, 2014

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge” arXiv:1409.0575, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, Advances in neural information processing systems, 2012

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks”, International Conference on Learning Representations (ICLR 2014), 2014.

[10] M. Oquab, L. Bottou, I. Laptev and J. Sivic, “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks”, Conference in Computer Vision (CVPR), 2014.

[11] K. Simonyan, A. Vedaldi and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014