CNN을 이용한 소셜 이미지 자동 태깅

Automatic Tagging for Social Images using Convolution Neural Networks

저자
(Authors)

장현웅, 조수선
Hyunwoong Jang, Soosun Cho

출처
(Source)

정보과학회논문지 43(1), 2016.01, 47-53 (7 pages)
Journal of KIISE 43(1), 2016.01, 47-53 (7 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE06585735

APA Style

장현웅, 조수선 (2016). CNN을 이용한 소셜 이미지 자동 태깅. 정보과학회논문지, 43(1), 47-53.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 43, No. 1, pp. 47-53, 2016. 1
http://dx.doi.org/10.5626/JOK.2016.43.1.47

CNN을 이용한 소셜 이미지 자동 태깅
(Automatic Tagging for Social Images using Convolution Neural Networks)

장 현 웅 [†]　　조 수 선 [††]
(Hyunwoong Jang)　　(Soosun Cho)

요 약 인터넷이 급속히 발달하는 가운데 스마트폰, 디지털 카메라, 블랙박스 등의 기기에서 수집되는 방대한 영상 데이터가 소셜 미디어 사이트를 통해 빠르게 공유되고 있다. 소셜 미디어 공유 사이트에서는 일반적으로 이미지의 태그 정보를 사용하는데, 멀티미디어를 공유하는 방법이 쉬워지고 그 양이 폭발적으로 증가함에 따라 이미지에 태그를 붙여야 하는 일은 번거로움이 되고 있다. 또한 태그가 잘못 붙여지거나 안 붙은 경우에는 이미지 검색 정확도가 떨어질 가능성이 있다. 본 논문에서는 이미지의 내용정보를 이용하여 자동으로 이미지로부터 태그를 추출하는 방법을 제안한다. 제안하는 방법은 ImageNet에서 제공하는 대용량의 이미지 데이터와 라벨을 CNN(Convolutional Neural Network) 딥러닝 기법으로 학습시킨 후, 인스타그램 이미지로부터 라벨 정보를 추출하는 것이다. 추출된 라벨 정보를 이용하여 자동 태깅한 후, 검색에 활용했을 때 인스타그램의 기존 검색보다 높은 정확도를 가지고 있음을 알 수 있었다.
키워드: 이미지 자동 태깅, convolutional neural network, 이미지 내용정보, 인스타그램

Abstract While the Internet develops rapidly, a huge amount of image data collected from smart phones, digital cameras and black boxes are being shared through social media sites. Generally, social images are handled by tagging them with information. Due to the ease of sharing multimedia and the explosive increase in the amount of tag information, it may be considered too much hassle by some users to put the tags on images. Image retrieval is likely to be less accurate when tags are absent or mislabeled. In this paper, we suggest a method of extracting tags from social images by using image content. In this method, CNN(Convolutional Neural Network) is trained using ImageNet images with labels in the training set, and it extracts labels from instagram images. We use the extracted labels for automatic image tagging. The experimental results show that the accuracy is higher than that of instagram retrievals.
Keywords: automatic image tagging, convolutional neural network, image content, instagram

· 2015년 한국교통대학교 지원을 받아 수행하였음 또한, 이 논문은 2010년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2010-0013307)

† 학생회원 : 한국교통대학교 컴퓨터정보공학과
jhwskorg@gmail.com
†† 종신회원 : 한국교통대학교 컴퓨터정보공학과 교수
(Korea National University of Transportation)
sscho@ut.ac.kr
(Corresponding author임)

논문접수 : 2015년 5월 11일
(Received 11 May 2015)
논문수정 : 2015년 9월 3일
(Revised 3 September 2015)
심사완료 : 2015년 10월 28일
(Accepted 28 October 2015)

1. 서 론

인터넷이 고도로 발달하고 다양한 기기를 통해 영상 데이터를 공유하면서 플리커, 페이스북, 인스타그램과 같은 소셜 미디어 공유 사이트가 급격히 성장하고 있다. 일반적으로 소셜 미디어 공유 사이트에서는 대량의 영상 데이터를 관리하기 위해 사용자에 의한 태깅 방식을 사용한다. 전통적으로 웹 이미지 검색은 이미지에 달린 태그를 기반으로 하였다. 그러나 일반적으로 폭소노미 기반의 웹 이미지에 달린 태그는 각 개인의 주관적인 판단에 의한 것이기 때문에 이미지 원래 정보와는 다른 잡음 태그를 포함하고 있다[1-3]. 또 최근에는 다양한 기기를 통해 이미지가 공유되면서 태그가 없는 경우가 자주 발생하고 있다. 따라서 방대한 양의 이미지 데이터가 웹 공간에 빠르게 저장되는 환경에서 효율적인 이미지 검색이 더욱 중요해지고 있으며 이와 함께 정확한 태깅의 필요성이 부각되고 있다.

한편, 이미지의 내용을 분석하여 활용하고자 하는 요구가 증대되어 지난 십여 년 동안 내용기반 이미지 검색 방법(Content-based Image Retrieval)에 대한 연구가 다양하게 진행되어 왔다[1,4]. 여기서는 시각 단어집(Bag of Visual Words : BoVW)을 기반으로 이미지를 표현하고 분류하는 방법이 자주 사용되어 왔다[4-6]. BoVW 방법은 먼저 SIFT(Scale Invariant Feature Transform)나 SURF(Speeded Up Robust Features)와 같은 특징점 추출 알고리즘을 통해 이미지에서 변화에 강인한 특징을 추출한 후, 추출된 특징들로 군집화 과정을 수행함으로써 단어집(BoVW)를 구성한다. 구성된 시각 단어집으로 각 이미지의 내용정보를 표현할 수 있는데, 표현된 정보를 텍스트로 매핑시키면서 자동으로 이미지에 대한 설명을 추가하는 어노테이션에 사용할 수 있다. 이 때, 추출된 시멘틱 텍스트를 태그로 표현하여 해당 이미지에 자동 태깅함으로써 더욱 효과적이고 정확한 검색이 가능하게 되었다[5]. 그러나 시각 단어집 구성에 사용되는 SIFT 또는 SURF 알고리즘은 물체(object) 인식 분야에 뛰어난 성능을 보이지만, 배경(scene) 이미지 분류에서 정확도가 떨어진다는 단점이 있다[6].

최근에는 딥러닝(Deep learning)을 사용한 이미지 인식 방법인 CNN(Convolutional Neural Network)이 이미지 인식분야에서 뛰어난 성능을 보여 각광을 받고 있다[7,8]. 딥러닝 기반의 CNN은 웹 이미지의 폭발적인 증가와 GPU와 같은 컴퓨팅 시스템의 발전으로 기계학습(Machine Learning)분야에서 좋은 성능을 보이고 있다. CNN은 이미지에 컨볼루션 필터를 사용하여 계산이 빠르고, 특정 객체뿐만 아니라 배경을 포함한 이미지 전체를 고려할 수 있다는 점이 특징이다.

본 연구에서는 딥러닝 기반의 인식기법인 CNN을 사용하여 이미지에 자동으로 태그를 붙임으로써 이미지 전체를 고려한 풍부한 내용의 태깅이 가능함을 보이고자 한다. 특히 테스트 이미지 데이터 셋으로는 페이스북이 인수하면서 최근 실사용자 수가 트위터를 넘어선 이미지 기반의 소셜 미디어 공유 사이트인 인스타그램[9]을 사용함으로써 실험용으로 정제된 이미지가 아닌 스마트 기기를 통해 게시된 사용자들의 일상 사진들에서 자동 태깅이 얼마나 효과적인지 알아보고자 한다. 이 방법으로 기존의 SIFT기반 BoVW기법보다는 물론이고, 사용자가 직접 선택하는 태그들보다도 더 다양하고 풍부한 태깅이 가능함을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2절에서는 이미지 자동 어노테이션 기술 및 CNN 이미지 내용분석과 관련된 기존 연구에 대하여 설명한다. 3절에서는 본 논문에서 제안한 방법인 CNN을 사용해서 자동 태깅하는 방법과 실험을 소개한다. 4절에서는 결과를 분석하고 평가한다. 마지막으로 5절에서는 결론 및 향후 연구를 제시하며 마무리한다.

2. 관련 연구

2.1 이미지 내용기반의 자동 시멘틱 어노테이션

내용기반 이미지 검색에서 BoVW(Bag of Visual Words)기반의 분류 기법이 좋은 성능을 보이고 있다[6,10]. BoVW를 구성하기 위해 사용되는 특징점 추출 알고리즘인 SIFT는 이미지 인식에 주로 사용되는 방법으로 이미지의 색상, 위치, 크기, 회전 등의 변화에 강인한 특징점을 추출한다. SIFT알고리즘은 4×4 배열에 8개의 방향으로 나누어진 벡터들을 합한 128개의 dimension을 갖도록 Keypoint descriptor와 특징점의 위치 좌표(x, y), scale, 및 orientation의 정보를 가진 frame을 구성한다.

이미지로부터 특징점을 추출한 후에 k-means 알고리즘을 사용해서 k개의 visual words가 되는 local visual blocks, 곧 BoVW를 얻을 수 있다. 어노테이션을 위하여 local visual blocks 집합 R_i와 텍스트 집합으로 구성된 주석 keyword K_j 사이의 조건부 확률 $P(K_j|R_i)$을 통해 매칭 등급이 높은 순서대로 정렬하여 적합한 주석을 얻을 수 있다.

하지만 SIFT나 SURF와 같은 알고리즘은 Natural Scene Retrieval과 같이 영상의 특징적인 요소가 떨어지는 배경이미지에서는 좋은 성능을 보이지 못했다[6]. 최근에는 영상 인식 분야에서 딥러닝(deep learning)기법을 사용한 CNN(Convolutional Neural Network) 방법이 컨볼루션 영상처리 방법을 사용해서 SIFT가 가지고 있던 단점을 극복하여 이미지 인식에 뛰어난 성능을

보이고 있다[11].

SIFT기반의 BoVW는 이미지로부터 특징점을 추출한 후에 local visual blocks, 즉 BoVW를 구성하는 방법으로서, 기본적으로 미분을 이용한 이미지 화소값의 변화 정도를 측정하여 그 변화가 큰 영역들을 뽑아내어 특징 벡터를 구성하는 것이므로 배경 이미지보다는 뚜렷한 경계를 지닌 객체 이미지에서 더 효과적이다. 반면 CNN이 기초로 하는 딥러닝 기법은 별도의 특징 추출 알고리즘 없이 패턴을 인식하고 분류할 수 있다. 이렇게 하나의 신경망 네트워크에서 특징 추출과 인식이 통합적으로 이루어지기 때문에 계층들의 연관성을 활용하여 배경 이미지에서도 좋은 결과를 얻을 수 있다.

2.2 딥러닝 기법을 사용한 CNN 이미지 내용분석 방법

딥러닝(Deep learning)은 신경망 네트워크로 많은 수의 계층을 만들어 학습을 하는 기계학습 분야이다. 신경망 네트워크는 오래 전인 1950년대부터 인간의 뇌에서 영감을 얻어 시작되었지만 낮은 컴퓨팅 성능 문제와 XOR 연산 문제 등으로 인해 잠시 사라졌다가 최근 컴퓨팅 성능, 빅데이터, RBM(Restricted Boltzmann machine)으로 overfitting 문제해결 등을 통해 다시 부활하였다[8,12]. 보통 프로그램은 1차원 연산을 수행하는데, 신경망 네트워크는 병렬 연산을 수행하여 빠르고 복잡한 구조를 갖는다. 일반적으로 각 계층에는 여러 개의 노드들이 있는데, 입력 노드에서 다음 노드로 연결될 때 가중치 연산을 통해 두 노드 사이의 신호를 제어하여 하나의 값으로 출력한다. 일반적으로 퍼셉트론(perceptron)의 예를 들 수 있는데, l번째 계층의 노드 i에서 출력값 $Z_i^{(l)}$은 $l-1$번째 계층의 출력 값들과 가중치 $w_{i,k}^{(l)}$, bias $b_i^{(l)}$의 연산을 통해 식 (1)로 나타낼 수 있다.

$$Z_i^{(l)} = \sum_{k=1} w_{i,k}^{(l)} y_k^{(l-1)} + b_i^{(l)} \quad (1)$$

입력단계에서 들어온 값이 가중치 학습을 통해 조금 더 높은 수준의 특징을 갖게 되고, 이것이 여러 계층에 쌓이면서 상위 계층으로 갈수록 높은 수준의 특징을 추출하게 된다. 신경망 네트워크에서는 에러(예측한 값과 실제 값의 차이)를 줄이기 위해 지속적으로 가중치 값을 학습한다. 이러한 학습은 오류 역전파 알고리즘(error backpropagation algorithm)을 통해 이루어지는데, 그 방법으로 경사도 연산 방법(Computing the Gradients)을 주로 사용한다[8]. 신경망 네트워크에서 에러를 E(w)로 표기하는 cost function이라고 할 수 있고, 식 (2)로 나타낼 수 있다. 여기서 $f(Z_i^{(l)})$은 퍼셉트론에서 연산을 하고 출력된 값에 활성화 함수를 적용한 예측 값이다. a는 실제 목표 값으로 식 (2)는 신경망을 통해 나온 예측 값과 목표인 실제 값의 차이를 알아보기 위한 연산이다.

$$E(w) = \frac{1}{2}\sum_{i=1}^{N} \| f(Z_i^{(l)}) - a \|^2 \quad (2)$$

에러의 최적화를 통해 신경망 네트워크를 학습시키는 것이다. E(w)을 경사도 연상 방법으로 최적화시키는 식 (3)를 만들 수 있다. E(w)를 미분해서 변화도$\frac{\partial E}{\partial w_{ij}}$를 찾은 후 최솟값이 될 수 있도록 가중치 값을 이동시키는 것을 반복한다.

$$w_{ij} \leftarrow w_{ij} + \frac{\partial E}{\partial w_{ij}} \quad (3)$$

이러한 신경망 네트워크의 특징은 다른 인식 알고리즘들과 다르게 특징 추출과 인식이 하나의 신경망을 통해서 이루어진다는 점이다. 기존에는 특징 추출 알고리즘을 통해 특징 벡터를 추출한 후, 벡터 기계학습을 통해 분류나 인식을 하지만 딥러닝은 별도의 특징 추출 알고리즘 없이 패턴을 인식하고 분류할 수 있다. 이렇게 하나의 신경망 네트워크에서 특징 추출과 인식이 통합적으로 이루어지기 때문에 계층들의 연관성을 활용하여 좋은 결과를 얻을 수 있다.

특히 토론토 대학에서는 대규모 이미지 검색을 위해 CNN(Convolutional Neural Network) 기술을 개발하여 Caltech 이미지 인식, ImageNet 이미지 인식 등에서 기존보다 뛰어난 기록을 세우고 있다[13]. CNN의 구조는 그림 1과 같이 크게 Convolution 계층, Pooling 계층, Fully Connected 계층으로 이루어진다.

CNN에서는 하위 계층부터 상위 계층을 지나면서 점차 수준이 높은 특징을 추출한다. 하위 계층에서는 복수의 convolution과 pooling을 통해 특징맵(feature map)을 구성한다. Convolution 계층에서는 이전 계층의 복수의 출력 값을 입력받아 공유된 가중치 연산(convolution filters)처리를 수행하고, pooling 계층에서는 convolution 계층과 1:1로 연결되어 max-pooling을 수행한다. Max-pooling에서는 블록 내의 특징값 중 최대값을 취함으로써 위치에 상관없이 특징이 되는 값은 보존하고 특징맵의 크기를 줄여 연산을 빠르게 해주는 역할을 한다. 최상위 fully-connected 계층에서는 이전 계층에서 추출된 높은 수준의 특징을 사용해 최종 인식 결과를 결정하게 된다.

최근 CNN 기법을 사용하여 이미지에 의미 정보를 자동으로 태깅하기 위한 다양한 연구들이 진행되고 있다[12,14]. 이들은 연구[13]의 기본적인 CNN구조를 사용하여 다양한 방법으로 이미지 분석을 시도하고 있는데, 이를 위해 미리 수집된 정제된 이미지 데이터 셋으로 CNN을 훈련시킨다. 일반적으로 PASCAL Visual Object Classes, MIT Indoor Scene, ImgaeNet, Caltech 101 등 컴퓨터 비전 분야의 실험을 위한 정제된 이미지

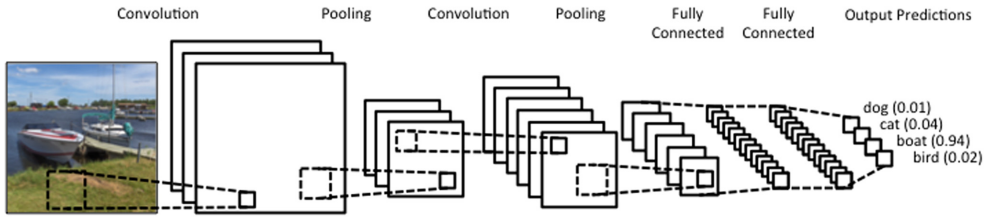


그림 1 Convolutional Neural Network 구조
Fig. 1 The architecture of CNN

데이터 셋이 사용되고 있다. 이들 연구에서는 이미지에 라벨이 붙여진 데이터 셋으로 CNN을 훈련시킨 후 동일 데이터 셋에서 분류 실험을 한다. 연구[14]의 실험에서는 훈련과 분류에 사용된 데이터 셋에 따라 다른 결과를 보이고 있다. PASCAL의 이미지 셋으로 실험한 결과는 평균 72.6%의 이미지 자동 태깅 정확도를 보였고, MIT Indoor 이미지 셋에서는 평균 59.6%의 정확도를 보이고 있다. 이것은 기존의 태깅 방법에 비해 확실히 더 나은 결과임이 분명하다[12,14]. 위 실험들에서 테스트 이미지들이 훈련 이미지 셋에서 무작위로 추출된 것이라 해도 인위적으로 수집되고 정제된 데이터이므로 실제 급변하는 인터넷 공유사이트에 사용자들에 의해 저장된 이미지들을 대상으로 하면 어떤 결과가 얻어질 것인지 확인할 필요가 있었다. 본 연구에서는 CNN기법을 구현하여 현재 소셜 이미지 사이트로 가장 각광 받고 있는 인스타그램 이미지들을 대상으로 자동 태깅 및 분류실험을 수행하였다.

3. 구현 및 실험

CNN을 구성하기 위해 토론토대학의 Network모델을 사용하였다[15]. CNN을 훈련하기 위한 이미지로는 ImageNet 2012를 사용하였는데, ImageNet 데이터 셋은 WordNet 계층의 명사들에 따른 셋으로 약 120만개의 이미지들과 1,000개의 클래스 카테고리로 구성되어 연구용으로 제공되고 있다[13].

또, 최근 페이스북이 인수하여 데이터가 폭발적으로 증가하고 있는 인스타그램의 이미지를 테스트 데이터 셋으로 사용했다. 인스타그램은 사용자가 일상생활 사건을 스마트 기기를 사용해서 촬영하고 게시하는 멀티미디어 공유 사이트로 이를 활용한 다양한 연구가 진행되고 있다[5,9]. 인스타그램을 선택한 이유는, 일상생활에서 사용자들이 업로드하는 이미지이기 때문에 정제되지 않았고 사용자의 주관적인 태깅으로 태그기반의 정확한 검색이 쉽지 않기 때문이다. 본 실험에서는 이미지 검색의 정확도를 증가시킬 수 있다는 것을 증명하기 위해 총 450개의 인스타그램의 이미지 데이터 셋을 사용해서

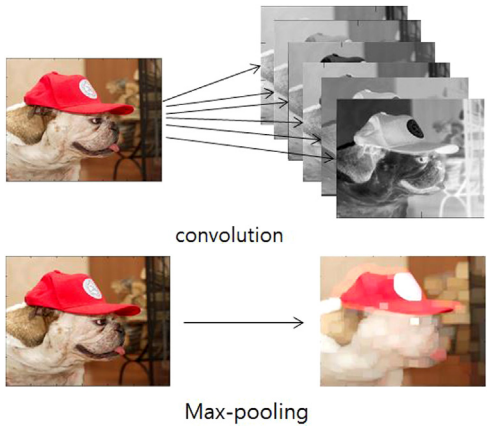


그림 2 convolution, Max-pooling 과정
Fig. 2 Processing of convolution and Max-pooling

15개의 카테고리 별로 30개씩의 이미지들을 테스트하였다.

CNN은 전체적으로 5개의 convolution, max-pooling, normalization 계층과 3개의 fully-connected 계층으로 구성했다. 그림 2와 같이 convolution 계층에서 여러 번의 convolution, max-pooling, normalization을 수행하면서 이미지의 정보들이 겹치게 되는데, 이로써 이미지의 특징을 불변하도록 학습할 수 있고, SIFT나 SURF와는 달리 전체적인 이미지 특징을 고려하기 때문에 배경이미지에서도 강한 성능을 보이고 있다.

N×N의 입력 이미지가 convolution 계층에 연결되어 있고, m×m의 convolution 필터가 있을 때 convolution layer의 크기는 (N-m+1)×(N-m+1)이 된다. 각 convolution 계층은 이전 계층의 출력으로부터 $y_{ij}^{(l)}$를 입력을 받는다.

$$\text{for } i,j = 0,1,\ldots,N-m \tag{4}$$

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

$$y_{ij}^{(l)} = f(Z_{ij}^{(l)}),\ f \text{ is nonlinear function} \tag{5}$$

$$f(x) = \max(0,x),\ f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & otherwise \end{cases} \tag{6}$$

식 (4)에서는 이전 계층의 입력 값들을 받아 convolution 필터로 가중치를 계산한다. 식 (5)는 식 (4)에서 출력된 값을 활성화 $f(Z_i^{(l)})$ 함수로 계산한 값이다. 활성화 함수는 식 (6)에서와 같이 ReLU(Rectified Linear Unit)함수를 사용한다. 기존에 sigmoid함수를 사용하는 대신에 ReLU를 사용하면서 계산량이 줄었고 정확도는 증가했다. ReLU는 0이상인 값은 출력해주고, 0이하인 값들을 0으로 만들어 계산량을 줄일 수 있었다.

정확하고 의미 있는 특징을 추출하기 위해서 weight인 convolution 필터를 훈련시키는데, 이것은 오류 역전파 알고리즘(Backpropagation)과 경사도 연산 방법을 통해 이루어진다. 그림 3과 같이 최상단의 3개의 계층에서 fully-connected를 수행하게 된다. FC6와 FC7 계층에서지속적인 훈련과 특징을 추출하고, 마지막 계층에서는 최종적으로 softmax 분류를 수행하게 된다.

4. 결과분석 및 평가

표 1은 인스타그램에서 검색한 이미지에 달린 태그들과 같은 이미지를 CNN을 통해 분석하고 태그를 붙인 결과를 비교한 것이다. CNN에 의한 자동 태깅 결과, 기존에 붙은 태그보다 더 섬세하고 정확한 태그가 붙여진 것을 볼 수 있다.

표 2는 각 카테고리별로 30개의 이미지를 태그기반으로 검색한 후 그 정확도를 조사한 결과이다. CNN을 사용해서 다시 태그를 붙인 결과 총 15개의 카테고리에서 대부분 더 정확한 결과를 얻을 수 있었다. 전체적인 합계를 보면 기존 방법인 SIFT기반의 BoVW기법으로 붙여진 태그로는 36.67%, 인스타그램의 기존 태그로는

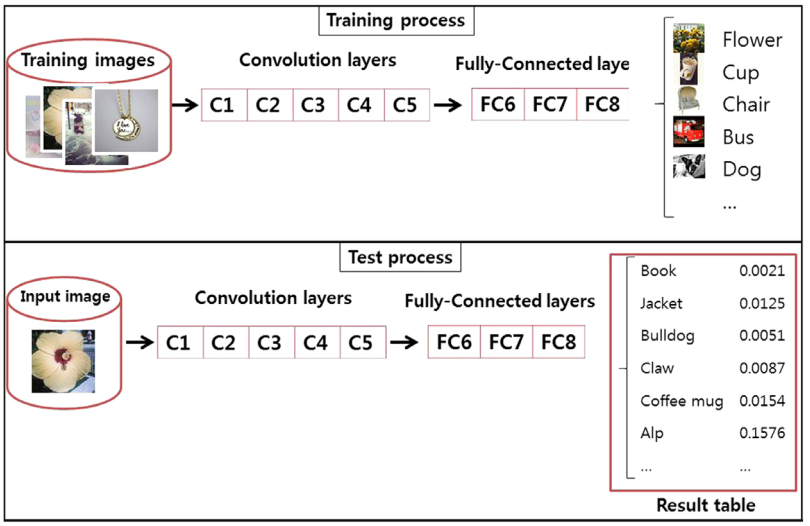


그림 3 전체적인 CNN 구조
Fig. 3 The architecture of CNN

표 1 인스타그램 태그와 CNN을 통해 태깅한 결과 비교
Table 1 A comparison of tagging results using CNN and instagram tags

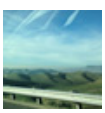
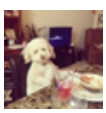
Image	Instagram Manual tags	Automatic tags using CNN	Image	Instagram Manual tags	Automatic tags using CNN	Image	Instagram Manual tags	Automatic tags using CNN
	bike smoke pretty	chain necklace bottlecap		bird photography ocean sun	albatross mollymawk pelican goose		bottle nice style	stole wool prayer rug prayer mat
	car accordion travel	wing valley vale alp		chair dog design pet	miniature poodle standard poodle toy poodle		house office luxury terrace	patio terrace home theater monitor

표 2 CNN을 사용해서 태깅한 결과의 검색 정확도 향상
Table 2 Improvement of retrieval correctness from tagging results using CNN

Categories	SIFT		instagram		CNN	
	correct tags	Rates (%)	correct tags	Rates (%)	correct tags	Rates (%)
bike	3	10.00	20	66.67	23	76.67
bird	20	66.67	15	50.00	21	70.00
bottle	15	50.00	23	76.67	21	70.00
building	3	10.00	23	76.67	15	50.00
bus	6	20.00	8	26.67	21	70.00
car	19	63.33	23	76.67	28	93.33
chair	5	16.67	19	63.33	18	60.00
cup	18	60.00	19	63.33	27	90.00
flower	19	63.33	18	60.00	18	60.00
food	6	20.00	12	40.00	22	73.33
house	10	33.33	7	23.33	17	56.67
mountain	3	10.00	17	56.67	21	70.00
pet	15	50.00	24	80.00	22	73.33
roadsign	22	73.33	27	90.00	22	73.33
sea	1	3.33	19	63.33	21	70.00
Total	165	36.67	274	60.89	317	70.44

60.89%의 검색 정확도가 얻어졌다. 반면 본 연구에서 제안한 방법의 결과는 70.44%로 나타나 기존 방법인 SIFT기반의 BoVW뿐만 아니라 인스타그램 고유의 태그를 이용한 검색보다도 약 10% 정확도가 증가한 것을 알 수 있다.

5. 결론 및 향후 연구

본 연구에서는 최근에 각광을 받고 있는 딥러닝 기반의 이미지 인식 방법인 CNN을 이용하여 소셜 미디어 공유사이트의 이미지에 보다 정확한 태그를 붙여주기 위한 방법을 제안하고 구현하였으며, 실험을 통해 그 효과를 입증하였다. CNN은 이미지에 컨볼루션 필터를 사용하여 계산이 빠르고, 특정 객체뿐만 아니라 배경을 포함한 이미지 전체를 고려할 수 있다는 점에서 뛰어나다고 할 수 있다. 때문에 최근 여러 연구에서 CNN기반의 이미지 인식 및 분류를 위한 다양한 시도가 전개되고 있다. 이들 연구가 실험용 정제 이미지 데이터 셋을 사용하여 CNN의 성능을 입증한 것들인 반면, 본 연구에서는 최근 사용자수가 급증하고 있는 소셜 미디어 공유사이트인 인스타그램을 사용하여 사용자들이 직접 업로드하는 일상생활 이미지에서도 CNN기반의 자동 태깅이 효과적인지 검증하고자 하였다. 그 결과, 기존의 방법인 SIFT기반의 BoVW에 의한 태그는 물론이고, 사용자들이 직접 등록하는 태그들보다 더 높은 정확도와 섬세함을 가진 태그들을 얻을 수 있었다. 이번 실험에서 CNN기반의 자동 태깅을 이용하여 인스타그램 자체의 수동 태그들에 비해 약 10%의 검색 정확도를 향상시킴으로써 CNN에 의한 자동 태깅이 실제 대용량 이미지 공유 서비스에도 효과적으로 적용될 수 있음을 보였다.

빅데이터 시대에 진입하면서 방대한 양의 데이터를 사용해서 스스로 학습하고 분류하는 딥러닝을 사용한 연구가 지속적으로 좋은 성능을 보이고 있다. 영상인식 분야에서 CNN은 현재 인간의 판단수준을 따라잡기 위해 발전을 거듭하고 있다. 향후 연구에서도 CNN 관련 기술을 더욱 효과적으로 응용하고 발전시켜 나갈 계획이다.

References

[1] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions," *Proc. of the 2013 IEEE Conference on Computer Vision and Patten Recognition*, pp. 1618-1625, Jun. 2013.

[2] J. Cha, S. Cho, Y. Uh, S. Kim and H. Byun, "Image annotation using tag refinement," *Journal of KIISE : Software and Applications*, Vol. 39, No. 8, pp. 613-620, 2012. (in Korean)

[3] S. Lee and E. Hwang, "Image Retrieval Scheme using Spatial Similarity and Annotation," *Journal of KIISE : Databases*, Vol. 30, No. 2, pp. 134-144, 2003. (in Korean)

[4] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," *Workshop on Statistical Learning in Computer Vision, ECCV*, Vol. 1, No. 1-22, May. 2004.

[5] L. Meiyu, D. Junping, J. Yingmin, and S. Zengqi, "Image Semantic Description and Automatic Semantic Annotation," Control Automation and Systems

(ICCAS), 2010 International Conference on, pp. 1192-1195, 2010.
[6] H. Jang and S. Cho, "Image Classification Using Bag of Visual Words and Visual Saliency Model," *KIPS Trans. Software and Data Engineering*, Vol. 3, No. 12, pp. 547-552, 2014. (in Korean)
[7] M. Oquab, L. Bottou, I. Laptev and J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," *Computer Vision and Pattern Recognition(CVPR), 2014 IEEE Conference on*, pp. 1717-1724, 2014.
[8] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, 61, pp. 85-117, 2015.
[9] M. Nam, J. Kim and J. Shin, "A User motion Information Measurement Using Image and Text on Instagram-Based," *Journal of Korea Multimedia Society*, Vol. 17, No. 9, pp. 1125-1133, 2014. (in Korean)
[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.
[11] P. Fischer, A. Dosovitskiy and T. Brox, "Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT," *arXiv preprint arXiv:1405.5769*, May. 2014.
[12] J. Wu, Y. Yu, C. Huang and K. Yu, "Deep Multiple Instance Learning for Image Classification and Auto-Annotation," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3460-3469, 2015.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012.
[14] Y. Gong, Y. Jia, T. Leung, A. Toshev and S. Ioffe, "Deep Convolutional Ranking for Multilabel Image Annotation," *arXiv preprint arXiv:1312.4894*, 2013.
[15] G. E. Dahl, T. N. Sainath and G. E. Hinton, "Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8609-8613 May. 2013.

장 현 웅

2013년 한국교통대학교 컴퓨터정보공학과(학사). 2015년 한국교통대학교 컴퓨터정보공학과(석사). 관심분야는 데이터마이닝, 영상처리

조 수 선

1987년 서울대학교 계산통계학과(학사)
1989년 서울대학교 계산통계학과(석사)
2004년 충남대학교 컴퓨터과학과(박사)
1994년~2004년 한국전자통신연구원 소프트웨어연구소 선임연구원. 2004년~현재 한국교통대학교 컴퓨터정보공학과 교수. 관심분야는 데이터마이닝, 정보검색, 영상처리