다중 레이블 이미지를 활용한 CNN기반 이미지 어노테이션 시스템의 개선

김택수○, 김상범
네이버
taeksoo.kim@navercorp.com, sangbum.kim@navercorp.com

Improving a CNN-based Image Annotation System Using Multi-Labeled Images

Taeksoo Kim○, Sangbum Kim
Naver Corp.

요 약

최근 딥러닝 기술의 발전에 힘입어 이미지로부터 자동으로 관련된 단어 혹은 문장을 생성하는 연구들이 진행되고 있는데, 많은 연구들은 이미지와 단어가 1:1로 대응된 잘 정련된 학습 집합을 필요로 한다. 한편 스마트폰 보급의 확산으로 인스타그램, 폴라 등의 이미지 기반 SNS가 급속하게 성장함에 따라 인터넷에는 한 이미지의 복수개의 단어(태그)가 부착되어있는 데이터들이 폭증하고 있는 것이 현실이다.
본 논문에서는 소규모의 잘 정련된 학습 집합뿐 아니라 이러한 대규모의 다중 레이블 데이터를 같이 활용하여 이미지로부터 태그를 생성하는 개선된 CNN구조 및 학습알고리즘을 제안한다. 기존의 분류 기반 모델에 은닉층을 추가하고 새로운 학습 방법을 도입한 결과, 어노테이션 성능이 기존 모델보다 11% 이상 향상되었다.

주제어: Deep learning, Image annotation, Semantic embedding

1. 서론

모바일 기기의 확산과 소셜 네트워크 서비스(SNS)의 성장에 따라 수많은 사람들이 즉석에서 사진을 촬영하고, 태그나 주석을 달아 자신의 SNS 공간에 올리고 있다. 특히 인스타그램이나 폴라와 같은 이미지 기반 SNS 시스템에는 이러한 이미지-다중 레이블 형태의 데이터가 매 시간 폭발적으로 업로드되고 있다. 이미지-다중 레이블 형식의 데이터는 비교적 정제된 형태를 지니므로 태그 추천이나 태그 랭킹 알고리즘을 위한 학습 데이터로 사용 가능할 것으로 기대된다.

기존의 이미지 분석 관련 알고리즘은 대부분 단일 레이블의 이미지 데이터를 가정한다. 즉, 각 이미지에 하나의 정답 레이블이 주어진 데이터만을 학습 데이터로 받아들이게 된다. 그러나 단일 레이블 데이터는 대부분 수작업으로 만들어지므로 대량의 학습 집합을 구축하기 어렵다. SNS 시스템에 축적된 다중 레이블 데이터는 다소 노이지 하다는 단점이 있지만, 대량으로 존재한다는 점과, 폭넓은 어휘를 다루고 있다는 점에서 사용 가치를 지닌다. 뿐만 아니라 적절한 학습 방식을 도입한다면, 한 이미지에 함께 존재한 레이블들 간의 시멘틱 정보까지도 활용할 수 있다.

본 연구에서는 이미지 기반 SNS에 업로드 된 대량의 다중 레이블 이미지와, 수작업으로 만들어진 소량의 단일 레이블 이미지를 모두 활용한 이미지 어노테이션 시스템을 제안한다. 즉, 해당 시스템은 이미지들 입력으로 받고 이미지와 가장 관련된 단어들을 출력해준다. 본 연구에서 제안하는 모델은 기본적으로 이미지 분류 모델인 AlexNet[1] 의 변형체이며, 크게 2가지 요소로 특징지을 수 있다.

첫째, 기존 AlexNet의 분류층(단어노드층) 앞에 단어의 의미정보를 인코딩할 수 있는 새로운 은닉층을 추가하고, 입력 이미지에 대응되는 복수개의 단어 노드들에서의 평균손실함수[2]를 최소화시키는 방향으로 학습을 진행하였다. 즉, 이미지의 특징벡터(입력층)가 단어의 의미벡터(은닉층)를 거쳐 실제단어(출력층)로 연결되는 구조로 학습시키는데, 이를 통해 함께 출현하는 태그들의 의미적 유사성도 모델링될 수 있도록 하였다.

둘째, 레이블이 이미지를 잘 설명한다고 보장되는 단일 레이블 데이터와 이미지와의 관련성과, 태그간의 관련성을 동시에 부분적으로 갖는 다중 레이블 데이터는 그 성격이 다르기 때문에, 두 서로 다른 종류의 데이터를 학습 과정에서 다른 방식으로 사용하는 선별적 업데이트 방식을 제안한다.

학습을 위해 이미지 기반 SNS에 축적된 대량의 다중 레이블 이미지와, 직접 구축한 소량의 단일 레이블 이미지를 이용하였으며, 평가에는 학습에 사용되지 않은 단일 레이블 이미지를 사용하였다. 먼저 다중 레이블 데이터만 사용해 모델을 학습시킨 뒤 성능을 평가해보고, 소량의 단일 레이블 데이터를 이용해 모델을 파인튜닝 시킨 뒤 다시 성능을 평가하였다. 2장에서 간단히 관련연구들에 대해 살펴보고, 3장과 4장에서 제안하는 방법과

그 실험결과를 보이고자 한다.

2. 관련 연구

최근 급부상한 딥러닝 기술은 데이터에 내재된 구조적 특징을 스스로 학습한다는 점과, 특정 분야에 대한 노하우가 없이도 손쉽게 사용할 수 있다는 점으로 인해 큰 인기를 얻고 있다. 특히 이미지 분류 [1,3,4], 이미지 검출 [5,6] 등 컴퓨터 비전 분야에서 매우 가시적인 성과가 나타나고 있으며, 기계 번역[18,19]과 같은 자연어 처리 분야에서도 그 영역을 넓히고 있다.

딥 러닝을 이용해 이미지에서 텍스트를 추출하는 기술은 크게 분류기, 결합 임베딩, 이미지 캡셔닝의 세 종류로 나눌 수 있다. 먼저 분류기 방식[1,3,4]은 AlexNet과 같은 Convolutional Neural Network (CNN) 기반의 분류 모델을 의미하며, 주로 분류층의 노드 개수를 필요한 클래스의 개수만큼 늘려 사용한다. 두 번째는 결합 임베딩 [10,11] 으로, 이미지 특징 벡터와 텍스트 특징 벡터를 각각 추출한 뒤 두 벡터를 새로운 벡터 공간 내 인접한 공간으로 사상시키는 방식이다. 보통 이미지 특징 벡터 추출을 위해 CNN을, 텍스트 특징 벡터 추출을 위해 워드 임베딩[7,8,9]을 사용한다. 마지막으로, 이미지 캡셔닝 방식[12,13,14,15]은 CNN을 이용해 이미지 특징 벡터를 추출한 뒤 회귀 신경망 [16,17] 등을 이용해 재귀적으로 캡션을 생성해 낸다.

3. 다중 레이블 이미지 분석 시스템

본 논문에서는 그림 1에 나타난 기존의 AlexNet의 구조를 다중 레이블 이미지 분석에 적합한 구조로 바꾸고, 이에 따른 새로운 학습 방식을 제안한다.

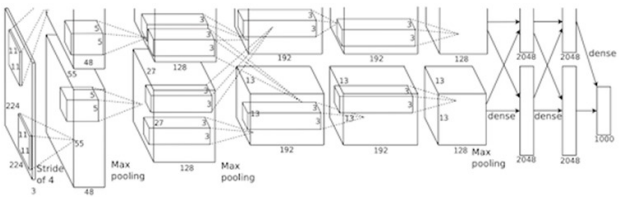


그림 1. AlexNet 구조도

AlexNet은 이미지를 입력으로 받아, 5개의 컨볼루션층(Conv1~Conv5)과 2개의 완전연결층(FC6, FC7)을 거친 뒤 마지막 분류층(OUT)에서 각 클래스에 대한 확률 값을 출력해준다. [1]에서는 OUT층에 1000개의 노드가 존재한다.

이 신경망을 학습시키기 위해서는 OUT층의 각 노드를 유일한 레이블로 갖는 학습이미지가 필요하다. 따라서 수만~수십만 개의 OUT층 노드로 확장하기 위해서는 마찬가지로 각 태그가 유일하게 할당된 이미지집합이 있어야 하는데 이를 수작업으로 구축하는 것은 매우 비용이 많이 든다. 따라서 본 논문에서는 이미 사진기반 SNS로부터 생성된 다소 부정확하나 대규모로 존재하는 다중 레이블 이미지 집합을 활용하도록 AlexNet을 그림 2과 같이 변형하였다.

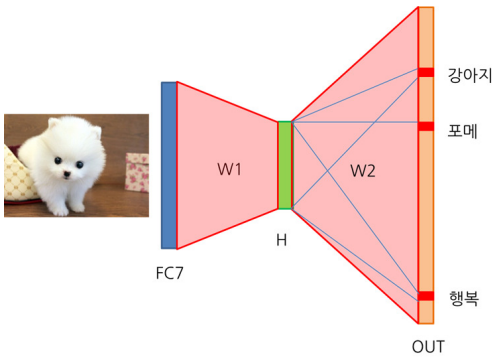


그림 2. AlexNet의 변형. FC7층과 OUT층 사이에 추가은닉층 H를 두었다

변형된 모델은 두 가지 특징을 지닌다. 첫 번째로, 그림 2과 같이 은닉층 H가 추가되었다. 이 추가은닉층은 이미지특징을 나타내는 FC7층과 OUT층 사이에서 시멘틱한 정보를 인코딩하는 역할을 하게 된다. 이 은닉층을 기준으로, 앞 부분의 W1에는 이미지의 특징으로부터 의미(개념)을 생성해내는 신경망이, 뒷 부분의 W2에는 의미(개념)으로부터 레이블(단어)을 생성해주는 신경망이 학습된다.

한편 학습을 위해서 [2]에서 사용한 평균 손실 함수를 도입하였다.

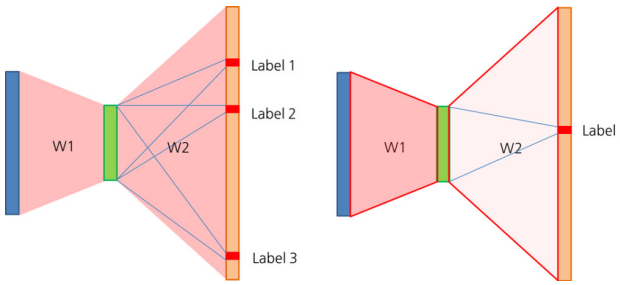
$$loss(i) = -\frac{1}{n_i}\sum_{t=1}^{n_i}\log P(w_{i,t}|V_i) \quad (1)$$

i는 데이터의 인덱스를 나타내며, n_i는 i번째 데이터에 부착된 레이블의 개수를 의미한다. V_i와 $w_{i,t}$와 는 각각 i번째 데이터의 이미지 특징벡터와, t번째 레이블을 의미한다. 확률함수 $P(x)$는 다음과 같이 소프트맥스 함수를 사용하였다.

$$P(x) = \frac{e^x}{\sum_k e^{x_k}} \quad (2)$$

두번째로, 다중 레이블 데이터와 단일 레이블 데이터를 서로 다른 방식으로 활용하는 선택적 부분 업데이트(SPU:Selective Partial Update) 방식이 사용되었다. 우선 다중 레이블 데이터를 사용해 학습을 진행할 때에는, 그림 3(a)와 같이 W1과 W2를 모두 학습(NU:Normal Update)시키게 된다. 이후 단일 레이블 데이터를 추가로 이용해 모델을 파인튜닝 할 때에는 그림 3(b)와 같이 W2를 그대로 두고 W1만 학습(PU:Partial Update)시킨다. 만약 단일 레이블을 (1)에 적용하면, 해당 단일 레이블의 확률을 최대화 시키고 나머지 모든 레이블의 확률은 최소화 시키는 방향으로 모델이 학습된다. 이는 다중 레이블 데이터에서 함께 출현한 레이블들의 확률을 동시에 높여 유사성을 인코딩한 것과 반대의 효과를 가져 온다.

즉, 레이블 간 유사성을 손실시키게 된다. 그러므로 단일 레이블을 이용할 때에는 레이블 간의 의미적 관계 정보는 유지하면서 이미지 특징으로부터 의미를 생성해주는 신경망만 학습시키는 PU 방식을 사용하였다.



(a) (b)

그림 3. 학습데이터 유형에 따른 선택적 부분업데이트 ((a) 다중 레이블 이미지를 이용한 NU, (b) 단일 레이블 이미지를 이용한 PU)

4. 실험 및 평가

4.1 데이터 및 평가척도

본 연구에서는 대량의 다중 레이블 이미지와 소량의 단일 레이블 이미지를 순차적으로 이용해 모델을 학습시켰으며, 학습에 사용되지 않은 단일 레벨 데이터를 이용해 성능을 평가하였다.

구체적으로 학습 및 평가에 사용된 데이터는 다음과 같다. 먼저 다중 레이블 데이터는 한 상용 이미지 기반 SNS에 2015년 2월부터 2015년 6월까지 업로드 된 약 40만개의 이미지-태그열 쌍으로 구성되며 모두 학습에만 사용하였다. 단일 레이블 데이터는 총 540개 클래스에 대해 수작업으로 구축하였으며, 이 중 380개 클래스에 대한 34,221개의 이미지를 파인튜닝용으로 사용하고, 540개 전체 태그에 대한 3,000개의 이미지를 추출해 평가용으로 사용하였다.[1)]

한편 평가 척도로는 다수의 관련연구들 [10,11] 과 마찬가지로 hit@k를 사용하였는데, 이 척도는 평가 이미지에 대한 상위 k개의 어노테이션 중 정답이 존재한 비율로 계산된다.

4.2 학습

모델의 추가은닉층은 1,000차원으로 구성하였으며, 마지막 분류층은 다중 레이블 데이터에 일정 빈도 이상 나타난 22,808개의 태그들로 구성하였다. AlexNet의 처음 7개의 층은 ImageNet의 Large Scale Visual Recognition Challenge 2012 데이터로 미리 학습된 값을 사용하였으며, 학습 과정에서 값을 변화시키지 않았다.[2)]

학습은 두 단계로 진행하였다. 먼저, 40만여개 다중 레이블 데이터로 모델을 학습하였다 (모델 A). 초기 학습 속도는 0.005, 미니 배치의 크기는 500으로 세팅했으며, 매 에포크마다 학습 속도를 5%씩 감소시켰다.

이후 단일 레이블을 이용해 파인튜닝할 때에는 그림 3의(b) 와 같이 부분업데이트를 적용했다 (모델 B). 초기 학습 속도는 0.001, 미니 배치의 크기는 500으로 세팅하고, 매 에포크마다 같은 방식으로 5%씩 학습 속도를 감소시켰다.

먼저 다중 레이블만으로 학습시킨 모델 A의 성능 측정을 위해, 각 이미지에 붙어 있는 첫 레이블만을 취해 학습한 경우 (first), 임의로 하나의 레이블을 선택해 학습한 경우 (random), 그리고 [2]와 같이 평균 학습 손실 함수를 적용한 경우(mean)의 세 가지 AlexNet을 각각 학습시켰다.

4.3 실험결과

표 1은 모델 A와 기존의 방식들을 비교한 결과이다. 전반적으로 제안된 모델이 3가지 AlexNet 기반의 모델보다 뛰어난 성능을 보이고 있으며, hit@k의 k값이 커질수록 점점 더 성능 향상의 폭이 커짐을 알 수 있다. 이는 본 연구에서 제안한 시멘틱-은닉층이 추가됨으로써 높은 빈도로 함께 나타났던 레이블들이 상위 결과에 함께 나타나는 효과로 보여진다.

다중 레이블과 단일 레이블을 모두 이용한 모델 B의 성능을 측정할 때에는 시멘틱-보존 업데이트 방식의 유효성을 측정하기 위해, 일반 업데이트만 사용한 비교모델(NU/NU)을 추가로 학습시켰다. 표 2에서 시멘틱-보존 업데이트 방식을 사용해 학습한 모델(NU/SPU) 이 다른 4개 모델에 비해 뛰어난 성능을 보임을 확인할 수 있다.

표 1. 다중 레이블만을 사용한 결과

	AlexNet (random)	AlexNet (mean)[2]	AlexNet (first)	Proposed
hit@1	0.048	0.046	0.048	**0.049**
hit@5	0.131	0.139	**0.146**	0.145
hit@10	0.191	0.203	0.212	**0.213**
hit@20	0,272	0.276	0.283	**0.298**
hit@30	0.320	0.328	0.328	**0.351**
hit@50	0.390	0.396	0.389	**0.429**

1) 평가에 사용된 레이블 중 30%가량은 학습 집합에 사용된 단일 레이블 데이터에는 존재하지 않도록 인위적으로 구성하여 실제 서비스 환경과 유사하게 하였다.

2) 미리 학습된 AlexNet 모델은 Caffe Model Zoo (http://caffe.berkeleyvision.org/model_zoo.html)의 BVLC CaffeNet을 사용하였다.

표 2. 다중 레이블과 단일 레이블 일부를 사용한 결과

	AlexNet (random)	AlexNet (mean)	AlexNet (first)	Proposed (NU/NU)	Proposed (NU/PU)
hit@1	0.134	0.145	0.136	0.142	**0.182**
hit@5	0.308	0.314	0.290	0.331	**0.374**
hit@10	0.383	0.401	0.366	0.410	**0.455**
hit@20	0.461	0.487	0.442	0.434	**0.545**
hit@30	0.507	0.535	0.487	0.555	**0.595**
hit@50	0.567	0.595	0.553	0.612	**0.662**

5. 결론

본 연구에서는 다중 레이블 이미지를 활용한 이미지 어노테이션 시스템을 제안하였으며, 기존의 분류 방식들과 비교했을 때 어노테이션 추출 성능이 최대 11% 이상 향상되었다.

본 연구의 의의는 다음과 같다. 먼저, 기존의 AlexNet 분류기에 은닉층을 추가하고 평균 손실 함수 기법을 도입해, 함께 나타난 레이블들 간의 유사성을 인코딩하는 방법을 제시하였다. 또한 선택적 부분 업데이트 방식을 도입하여 다중 레이블과 단일 레이블 이미지에 차등을 둠과 동시에, 학습된 시멘틱 정보를 유지하면서 성능을 상향시키는 학습 방식을 제시하였다.

향후에는 이미지-태그 형식의 데이터에 한정하지 않고 모든 형식의 이미지-텍스트 쌍을 훈련 집합으로 사용할 수 있는 모델에 대한 연구를 진행하려고 한다.

참고문헌

[1] A. Krizhervsky, I. Sutskever, G. Hinton, “ImageNet Classification with Deep Convolutional Neural Network”, Advances in Neural Information Processing Systems(NIPS), 2012.

[2] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, “Deep Convolutional Ranking for Multilabel Image Annotation”, arXiv, 2013.

[3] C. Szegedy, W. Kiu, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going Deeper with Convolutions”, Computer Vision and Pattern Recognition(CVPR), 2014.

[4] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv, 2014.

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

[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, “Overfeat: Integrated Recognition, Localization and Detection using Convolutional Networks”, arXiv, 2013.

[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space”, arXiv, 2013.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, Advances in Neural Information Processing Systems(NIPS), 2013.

[9] Q. V. Le, T. Mikolov, “Distributed Representations of Sentences and Documents”, arXiv, 2014.

[10] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, “DeViSE: A Deep Visual-Semantic Embedding Model”, Advances in Neural Information Processing Systems(NIPS), 2013.

[11] M. Norouzi, T.Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado , J. Dean, “Zero-Shot Learning by Convex Combination of Semantic Embeddings”, arXiv, 2013.

[12] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, “Show and Tell: A Neural Image Caption Generator”, arXiv, 2014.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, arXiv, 2014.

[14] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, G. Zweig, “From Captions to Visual Concepts and Back”, arXiv, 2014.

[15] A. Karpathy, L. Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions”, arXiv, 2014.

[16] S. Hochreiter, J. Schmidhuber, “Long Short-Term Memory”, Neural Computation, 1997.

[17] A. Graves, A. Mohamed, G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks”, IEEE International Conference on Acoustic, Speach and Signal Processing (ICASSP), 2013.

[18] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Learning”, arXiv, 2014.

[19] D. Bahadanau, K. Cho, Y. Bengio, “Neural Machine Translation by Jointly learning to Align and Translate”, arXiv, 2014.

[20] S. Venugopalan, H. Xu, J. Donahue, <. Rohrbach, R. Mooney, K. Saenko, “Translating Videos to

Natural Language Using Deep Recurrent Neural Network”, arXiv, 2015.