대규모 분류 체계에서 계층적 샘플링을 활용한 문서의 분류

홍성모○, 장헌석, 강인호

Naver Corporation
sungmo.hong@navercorp.com, heonseok.jang@navercorp.com, once.ihkang@navercorp.com

Classification using Hierarchical Sampling in Large Classification System

SungMo Hong○, HeonSeok Jang, Inho Kang
Naver Corporation

요 약

대규모 분류체계를 사용하는 경우, 기존 방법의 딥 러닝으로는 분류 정확도가 현저히 떨어진다. 이를 해결하기 위해 계층 구조를 활용한 네거티브 샘플링 방법을 제안한다. 학습 문서가 속한 카테고리의 상위 카테고리와 일정부분 겹치는 범위에서 네거티브 샘플을 선택하면, 하나의 큰 문제를 다수개의 하위 문제로 쪼개서 해결하는 학습 효과가 있다. 소규모 분류 체계와 대규모 분류체계 각각에서 샘플링 전략을 차용하였을 때를 비교한 결과, 대규모에서 효과가 좋았으며 그 때의 정확도가 150배 이상 차이가 나는 것을 보였다.

주제어: 문서 분류, 대규모 분류체계, 계층적 샘플링

1. 서론

데이터의 분류는 자료를 정보화하는 효율적인 수단이다. 하지만 데이터가 많아지면서 자연스럽게 데이터 분류 체계의 규모도 함께 늘어났다. 결국 증가한 데이터 개수 그리고 복잡한 분류체계로 인해 사람이 직접 데이터를 분류하기가 어려워졌다. 이를 해결하기 위해 토픽 모델링(Topic Modeling)에 대한 연구가 활발하게 진행되었다. 나이브 베이즈(Naïve Bayes)를 이용하여 글을 분류하거나, 잠재 디리클레 할당(Latent Dirichlet Allocation)을 사용하여 단어 분포를 가지고 문서의 주제를 예측하였다[1,2].

최근에는 딥 러닝을 이용하여 문서를 해석하고 분류하는 기술도 연구되었다. Yoon은 컨볼루션 신경망(Convolution Neural Network)를 사용하여 자연어 텍스트를 분류하는 방법을 연구하였다[3]. 이 방법은 소규모 분류체계에서 정확도가 높지만, 카테고리 개수가 많아질수록 정확도가 낮아진다. 이와는 달리 데이터의 유형에 따라 대규모 분류체계에서도 잘 동작하는 경우도 있다. 실제로 음성인식과 이미지 인식의 분야에서는 높은 정확도를 보여주는 모델 연구가 진행되었다[4,5].

자연어 처리 분야에서는 입력 문장을 다수개의 유형으로 분류한 연구결과가 있으나, 실험에 사용한 최대 분류 카테고리 개수는 6개이다[3]. 하지만 필요에 따라 대규모 분류체계를 사용할 수 있어야 하는데, 아직 이에 대한 연구는 많이 되지 않았다.

본 논문에서는 계층이 있는 대규모 분류체계 내에서 텍스트를 분류하고자 할 때, 학습에 사용하는 효과적인 샘플링 방법을 제시한다. 카테고리가 많은 특성 때문에 모델이 학습에 실패하는 현상을 극복하기 위해, 계층적 구조를 활용한 네거티브 샘플링(Negative Sampling)을 사용하여 학습을 국지적으로 수행하였다.

2. 관련 연구

잠재 디리클레 할당은 기 분류되어있던 문서의 단어 분포를 이용하여, 새로운 문서의 단어 분포를 보고 문서의 주제를 찾는다[1]. 하지만 이 방법은 단어 주머니(Bag-of-words) 방식이기 때문에, 단어가 속한 문맥을 정확하게 파악하는 데에는 어려움이 있다.

딥 러닝 기법인 컨볼루션 신경망과 순환 신경망(Recurrent Neural Network)은 단어가 속한 문맥을 포함하여 의미를 읽어낼 수 있다. Word2Vec은 흔히 볼 수 있는 텍스트를 학습 데이터로 사용하여 단어의 의미를 문장의 문맥에서 파악하고 임베딩한다[6]. 기계학습으로 텍스트의 의미를 파악하고 분류하는 연구도 진행되었다. Yoon은 입력 문장을 미리 정의한 카테고리들 중 하나로 할당하는 모델을 제시하였다. 이 모델의 구조는 Word2Vec, 컨볼루션 신경망, 최대값 풀링(Max Pooling), 그리고 완전 연결 신경망(Fully Connected Neural Network)을 함께 사용하여 만들었다. 하지만 본 논문의 실험 결과, 분류체계의 규모가 커지면 학습이 실패하여 분류 정확도가 줄어드는 것을 확인하였다.

하지만 음성 분야와 이미지 분야는 대규모 분류체계에서도 정확도가 높게 분류하고 있다. 음성인식 분야에서는 음성을 입력으로 순환 신경망을 사용하여 62개의 음소 카테고리 중 하나로 분류를 하였다[4]. 또한 이미지 분야에서는 컨볼루션 신경망을 사용하여 1000개의 카테고리 중 하나로 분류를 하였다[5].

본 논문은 Yoon이 제안한 모델의 구조를 변형하여, 완전 연결 신경망의 결과로 문서 벡터가 나오도록 학습한다. 문서 벡터와 가까운 카테고리 벡터를 주어진 문서의 카테고리로 할당할 수 있게끔 카테고리 벡터를 학습할 때, 카테고리 규모가 커서 학습이 되지 않는 문제점의

해결책을 제시한다.

3. 문서 분류를 위한 분류명 벡터 임베딩

동일한 벡터 공간에 분류명을 의미하는 벡터와 문서를 의미하는 벡터를 같이 표시할 수 있다면, 문서 벡터와 가까운 분류명 벡터를 찾아 문서의 분류명으로 예측하는 것이 가능하다. Zeynep은 벡터 공간에 입력 이미지에 대응하는 벡터와 이미지 레이블 벡터를 동시에 학습하여 주어진 이미지의 레이블을 예측하였다[7]. 이와 같이, 같은 공간에 다른 성격의 벡터를 함께 투사하는 것이 가능하다. 이를 텍스트 분류에 응용하여 문서 벡터와 분류명 벡터를 같은 공간에 투영하고, 최근접 이웃 탐색 알고리즘(K Nearest Neighbor)을 사용하여 주어진 문서를 분류하였다. 이처럼 벡터를 사용하는 이유는 계층적 분류체계에서 카테고리간의 거리가 다 똑같다고 규정할 수 없기 때문이다. 예를 들면, 동물 카테고리는 딥 러닝 카테고리보다 식물 카테고리와 거리가 가깝다고 할 수 있다. 이와 같은 이유로 각 카테고리를 벡터를 표현하여 카테고리간의 거리를 모델이 학습할 수 있게 구성한다.

Yoon의 모델은 점수 모델로 각 카테고리에 속할 상대적 확률을 출력한다. 학습 시에는 출력이 정답 카테고리의 확률만 1이 나오도록 확률적 경사 하강법(Stochastic Gradient Descent)으로 조절한다. 반면 본 논문에서 제시하는 모델인 벡터 모델은 문서를 의미하는 벡터를 출력한다. 문서 벡터와 정답 카테고리 벡터와의 유사도는 1이 되도록 그리고 오답 카테고리 벡터와의 유사도는 0이 되도록 학습한다. 본 논문에서의 유사도는 두 벡터 요소간 곱을 모두 합한 값을 사용하였다.

그림 1은 점수 모델과 벡터 모델의 학습과정을 설명하

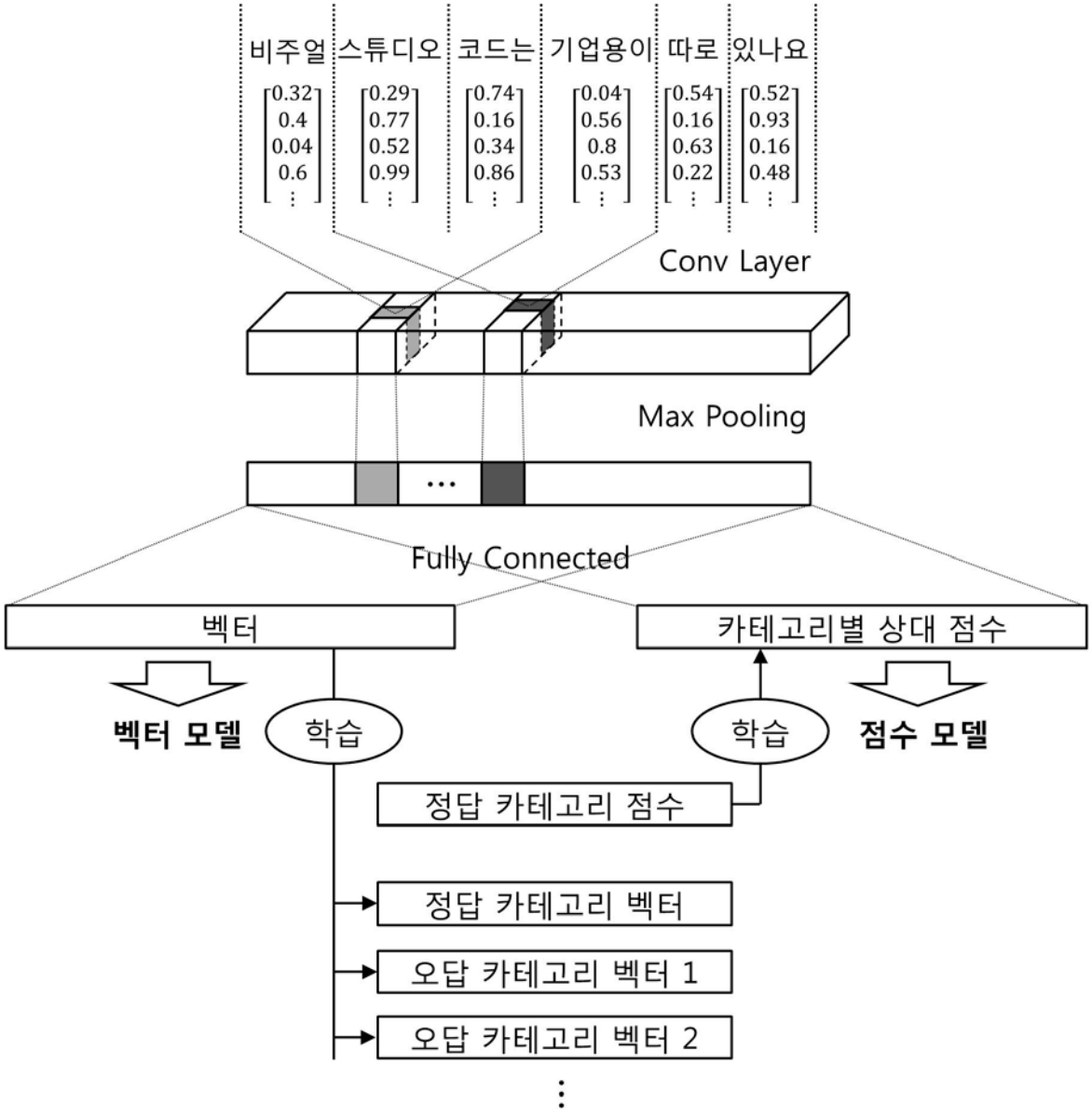

그림 1 벡터 모델과 점수 모델의 차이점 비교

고 그 차이를 보여준다. 텍스트를 컨볼루션 신경망 계층과 맥스 풀링(max pooling) 계층에 통과시키는 부분까지 동일하나, 완전 연결 계층을 거친 후의 결과를 해석하는 방식과 학습하는 방식이 다르다. 특히 벡터 모델이 문서 벡터의 결과를 사용해 카테고리 벡터를 학습시키는 점과, 다수개의 오답 카테고리도 학습에 영향을 받는다.

4. 대규모 분류체계에서의 학습 방법

분류체계의 규모가 큰 경우, 점수 모델을 사용해도 신경망의 가중치 값이 올바른 최적값에 도달하지 못하여 예측의 정확도가 현저히 낮아진다. 그 이유는 분류체계의 규모가 카테고리 벡터의 위치를 학습하기에 상대적으로 크기 때문이다.

이런 현상을 해결하기 위해 분류체계의 규모가 크면 분류가 계층적 구조를 차용하는 점을 이용한다. 계층적 구조의 특성에 따라 같은 분류명을 공유하는 하위 분류체계 또한 독립적인 분류체계의 특징을 가지고 있다. 그러므로 하위 분류 내에서 분류명 벡터 위치를 학습하는 하위문제(subproblem)들로 쪼개서 해결하면 전체 문제를 해결할 수 있다. 하위 문제를 효과적으로 해결하는 방법은 네거티브 샘플을 전략적으로 선택하는 것이다. 정답 카테고리의 상위 카테고리 중 하나의 카테고리를 대분류로 지정하여, 대분류 이하의 작은 분류체계를 학습하는 하위 문제로 재정의할 수 있다. 학습 데이터 한 개마다 다수개의 네거티브 샘플을 선택할 수 있기 때문에, 대분류를 다양하게 바꿔가며 고르게 표집하는 것이 효과적이다. 이와 같이 고른 샘플링 전략을 세우는 이유는 하나의 네거티브 샘플로 하위 문제를 해결함과 동시에 상위/하위 분류체계와도 연관성을 유지할 수 있기 때문이다.

그림 2는 샘플링 전략을 보여주는 구체적인 예시이다. 분류명 “라”에 속한 문서를 학습하고자 할 때, 각 숫자로 묶인 분류에서 임의로 하나의 분류명씩 네거티브 샘플로 선택하면 된다. 자세히 말해 학습 데이터의 분류가 상위 분류명부터 가-나-다-라인 경우, 첫 번째 샘플은 가-나-다에 속하면서 가-나-다-라에 속하지 않는 범위인 1번 그룹에서, 두 번째 샘플은 가-나에 속하면서 가-나-다의 모든 하위 분류에 속하지 않는 범위인 2번 그룹에서, 세 번째 샘플은 가에 속하면서 가-나의 모든 하위 분류에 속하지 않는 범위인 3번 그룹에서, 마지막 샘플은 가의 모든 하위 분류에 속하지 않는 범위에서 추출하는 것이다. 네거티브 샘플을 이용한 학습 방법은 이전 절에서 논의한 방법과 동일하다.

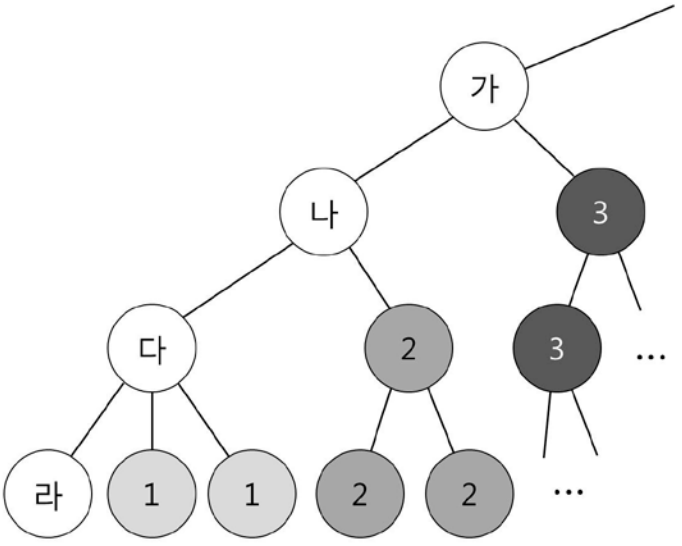

그림 2 계층적 구조에서의 네거티브 샘플링

5. 실험

5.1. 실험 데이터

네이버 지식인은 사용자 간 질의응답 플랫폼이다. 질문을 등록하는 분류체계의 규모가 크고 계층적 구조를 가지고 있다. 사용자가 지식인에 작성한 문서 중 800,852개를 학습에 사용하고, 19,624개를 개발용으로 사용, 39,249개를 평가에 사용하였다. 질문 데이터의 최상위 카테고리만 추출하여 13개 카테고리 규모의 소규모 분류체계를 만들고, 세번째 깊이와 그 상위 카테고리를 모아서 총 798개의 대규모 분류체계를 만들었다. 표 1은 분류체계 별 데이터의 특징을 보여준다.

표 1 분류체계별 데이터 특성

	소규모	대규모
카테고리 개수	13	798
단일 카테고리 내 최대 문서 수	122,261	10,881
단일 카테고리 내 최소 문서 수	4581	1
카테고리 내 평균 문서 수	61,604	1,003.5
표준편차	36954.5	1343.2
중앙값	58,920	836

5.2. 평가

평가 데이터의 문서를 사용, 가장 점수가 높은 분류명 N개에서 사용자가 선택한 분류명이 있는지 확인한다. 사용자가 선택한 분류명이 있으면 해당 예측은 정답으로, 없으면 오답으로 표시한다.

5.2.1. 소규모 분류체계

데이터에 등록된 분류명을 최상위 분류로 압축하면 총 13개의 분류로 모든 문서를 나눌 수 있다.

표 2와 그림 3은 소규모 분류체계에서 문서를 분류하였을 때, 상위 N개의 추천에서 적합한 분류명을 찾을 정답률이다. 분류명 수가 적을 때의 정답률은 벡터 모델에 비해 점수 모델이 미세하지만 소폭 나은 성능을 보였다. 이는 비계층적 형태의 분류체계이기 때문에 샘플링 전략

이 아무런 의미가 없었고, 카테고리 벡터간의 거리가 정확도에 영향을 미칠 만큼 다양하지 않기 때문으로 해석할 수 있다.

표 2 소규모 분류체계에서 N개 추천시 정답률

	벡터 모델	점수 모델
1개 추천	0.145	0.162
2개 추천	0.243	0.306
3개 추천	0.329	0.430

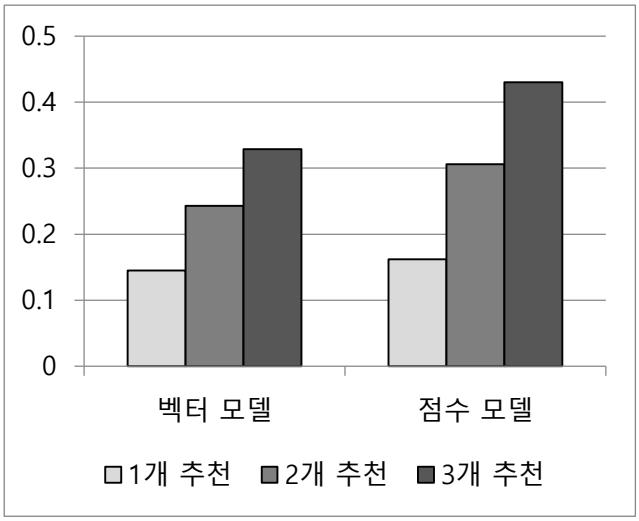

그림 4 소규모 분류체계에서 N개 추천시 정답률

5.2.2. 대규모 분류체계

데이터에 등록된 분류명을 깊이 3까지 압축하면 총 798개의 분류로 모든 문서를 나눌 수 있다. 마찬가지로 두 모델에서 점수가 높은 분류명 N개를 예측하고 정답률을 비교해 보았다.

표 3과 그림 4는 대규모 분류체계에서 문서를 분류하였을 때, 상위 N개의 추천에서 적합한 분류명을 찾을 정답률이다. 네거티브 샘플링(negative sampling)을 전체에서 무작위로 선발한 경우는 논문 제시 모델의 학습이 전혀 되지 않는 현상이 발견된다. 제시한 샘플링 원칙대로 선발할 경우, 논문 제시 모델의 성능은 좋아졌고 비교 모델의 성능은 나빠지는 경향을 보인다. 이는 벡터화 단계와 네거티브 샘플링 전략이 대규모 계층적 분류 체계 학습에 더 알맞다고 할 수 있다.

표 3 대규모 분류체계에서 N개 추천시 정답률

	벡터 모델	점수 모델
1개 추천	0.411	0.002
2개 추천	0.570	0.003
3개 추천	0.658	0.004

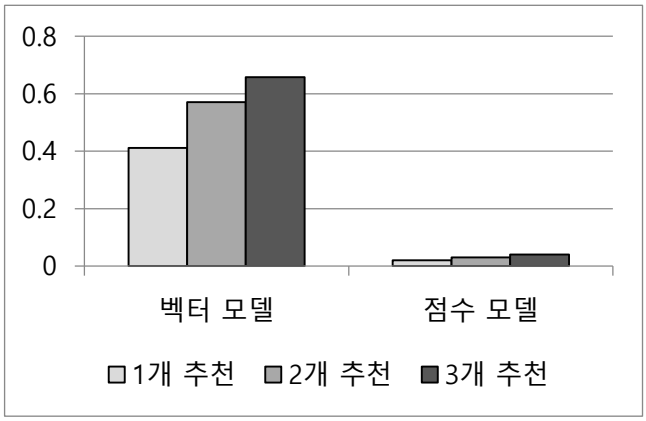

그림 3 대규모 분류체계에서 N개 추천시 정답률

표 4는 본 논문에서 제시한 모델로 일반 자연어를 분류했을 때, 잘 된 경우와 그렇지 않은 경우이다. 올바른 분류 판단 기준은 사용자의 질문과 사용자가 직접 등록한 카테고리를 비교하여 둘이 같은 경우에만 정답으로 판단한다. 첫 번째 오분석 예제의 경우 “ISP” 라는 단어를 정확하게 인지하지 못한 현상으로 보이며, 두 번째는 “몬츠”라는 의류 브랜드를 잘 학습하지 못한 현상으로 보인다.

표 4 분석과 오분석 예제

분석 예제	비주얼 스튜디오 코드는 기업용이 있나요? 회사 컴퓨터에 깔고 싶은데, 기업용이면 제한이 있어서..문의 드립니다.
	컴퓨터통신>프로그래밍
	리패키지?앨범 스밍 힘든건가요 1위하는게 어렵네여 ㅜ
	엔터테인먼트, 예술>음악>음악인
오분석 예제	isp 문자발송 관련건 질문 isp로 결제하게되면 카드명의 휴대폰으로 문자가 가나요?? 제 체크카드 명의가 부모님으로 되어있는데 문자 발송 가는건 별로 원하질 않아서욧....
	쇼핑>예약, 예매>여행상품
	몬츠 매장 ㅠㅠㅠ 몬츠 오프라인 매장 없나요? 작게 되있는 곳 말고 좀 크게 입점된 곳이요
	쇼핑>취미, 오락, 문구류>모형, 완구

6. 결론

분류체계의 규모가 크고 그 구조가 계층적일 때, 문서와 분류명을 임베딩하고 계층적 네거티브 샘플을 이용해 학습하여 적합한 분류명을 찾는 것이 일반적인 학습 방법에 비해 정확도가 150배 이상 높다. 반대로 분류 체계가 작고 단순한 경우에는 각 분류에 속한 문서의 주제가 방대해지기 때문에 하나의 카테고리 벡터로 표현하기 어렵다. 이러한 이유로 벡터 모델 방식은 대규모 분류 체계를 사용하는 경우에 오히려 성능이 낮아진다.

향후 사용자의 피드백을 받아, 기존의 학습된 벡터를 효과적으로 변경하는 방법에 대해 연구가 필요할 것으로 보인다.

참고문헌

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. Journal of machine learning research 3. Jan pp.993-1022, 2003.

[2] S. B. Kim, K. S. Han, H. C. Rim, and S. H. Myaeng, Some effective techniques for naïve bayes classification, IEEE transactions on knowledge and data engineering, 18(11), pp.1457-1466, 2006.

[3] Yoon Kim. Convolutional neural networks for sentence classification. In Proceedings of EMNLP. 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770-778, 2016.

[5] A. Graves, A. Abel-rahman Mohamed, and G. Hinton, Speech recognition with deep recurrent neural networks, IEEE international conference on acoustics, speech and signal processing, pp.6645-6649, 2013.

[6] Tomas Mikolov, et al. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

[7] Zeynep Akata, et al. Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence. 38.7: 1425-1438, 2016.