순환 신경망 기반 대용량 텍스트 데이터 분류 기술

조휘열[01] 김진화[2] 김경민[1] 장정호[4] 엄재홍[4] 장병탁[1,2,3]
서울대학교 공과대학 컴퓨터공학과[1]
서울대학교 인문대학 협동과정 인지과학전공[2]
서울대학교 자연대학 협동과정 뇌과학전공[3]
SK텔레콤 종합기술원[4]
{hyjo, jhkim, kmkim, btzhang}@bi.snu.ac.kr
{jeongho.chang, jaehong.eom}@sk.com[4]

Large-Scale Text Classification with Recurrent Neural Networks

Hwiyeol Jo[1] Jin-Hwa Kim[2] Kyung-Min Kim[1] Jeong-Ho Chang[4] Jae-Hong Eom[4] Byoung-Tak Zhang[1,2,3]
School of Computer Science & Engineering, Seoul National University[1]
Interdisciplinary Program in Cognitive Science, Seoul National University[2]
Interdisciplinary Program in Neuroscience, Seoul National University[3]
Corporate R&D Center, SK Telecom[4]

요 약

문서 분류 문제는 오랜 기간 동안 자연어 처리 분야에서 연구되어 왔다. 우리는 기존 컨볼루션 신경망을 이용했던 연구에서 더 나아가, 순환 신경망에 기반을 둔 문서 분류를 수행하였다. 순환 신경망에서는 가장 성능이 좋다고 알려져 있는 장기-단기 기억 (Long-Short Term Memory; LSTM) 신경망과 회로형 순환 유닛(Gated Recurrent Units; GRU)을 활용하였다. 실험 결과, 분류 정확도는 Multinomial Naive Bayesian Classifier, SVM, LSTM, CNN, GRU의 순서로 나타났다. 따라서 텍스트 문서 분류 문제는 시퀀스를 고려하는 것 보다는 문서의 feature를 뽑아 분류하는 문제에 가깝다는 것을 추측할 수 있었다. 그리고 GRU가 LSTM보다 문서의 feature 추출에 더 적합하다는 것을 알 수 있었으며 적절한 feature와 시퀀스 정보를 함께 활용할 때 가장 성능이 잘 나온다는 것을 확인할 수 있었다.

1. 서 론

텍스트 분류 문제는 자연어 처리 분야에서 오랜 기간 연구되어 왔으며, 그 응용분야도 다양하다. 예를 들어, 수많은 문장, 문서 속에서 원하는 데이터들을 걸러내는 필터링 (filtering)[1], 각 데이터가 어떤 정보를 담고 있는지에 대한 태깅 (tagging)[2]에 활용되는 등 처리할 데이터의 크기가 커짐에 따라 더욱 중요성이 강조되고 있다.

과거의 연구 방향이자, 지금도 뛰어난 성능으로 많이 사용되고 있는 Bag-of-Words[3] (BoW) 방법을 이용한 Naive Bayesian Classifier[4], 서포트 벡터 머신 (Support Vector Machine; SVM)[5] 등은 주로 단어의 빈도수를 feature로 사용하였기 때문에 해당 단어가 그 문장, 혹은 문단에서 어떤 의미로 쓰였는지 알기 힘들었다. 이를 극복하려는 시도로 한 번에 여러 단어를 보는 N-gram 방법을 통해 단어의 의미적, 문맥적 정보를 파악하려 했으나, 단순히 여러 단어를 보는 것만으로는 텍스트의 모든 의미를 파악하는데 한계가 있었다. 또한, N 값이 커질수록 계산 량이 급격히 늘어나는 단점도 있었다. 그러나 모든 문서 속 단어들의 의미적, 문맥적인 정보를 완벽하게 파악하지 않아도 적절한 성능이 나왔기 때문에 여전히 BoW 방법은 널리 사용되고 있다.

이전 연구에서 우리는 컨볼루션 신경망을 이용하여 문서들마다 커널을 통해 feature를 추출, 분류를 했다[6]. 그리고 커널의 크기를 변화시켜가며 N-gram을 모사하였다. 그 후속 연구로써 이번에는 언어 모델링, 음성 인식과 같은 순차적인 데이터에서 뛰어난 성능을 내고 있는 순환 신경망 (Recurrent Neural Networks; RNN)을 이용해 동일 데이터를 분류해보았다.

2. 분류 모델

문서 분류에 사용되는 대표적인 모델로는 Naive Bayesian Classifier, 서포트 벡터 머신, 컨볼루션 신경망, 순환 신경망이 있다.

2.1 Naive Bayesian Classifier

Naive Bayesian Classifier[4]는 각 사건들이 서로 독립이라는 가정을 한 후, Bayes' theorem을 이용하여 확률을 계산, 분류하는 모델이다. 따라서 두 확률의 결합 확률 (Joint Probability)을 두 확률의 곱으로 표현해버리지만, 상당히 강력한 성능을 보이고 있어서 널리 사용된다. 그러나 여전히 독립이 아닐 수 있는 두 사건을 독립으로 가정하기 때문에 한계가 존재한다.

Naive Bayesian classifier는 feature들 간의 조건부 독립 성질을 이용하는 반면, Multinomial Naive Bayesian Classifier는 feature들이 다항 분포 (multinomial distribution)를 따른다는 정보를 활용한다.

2.2 서포트 벡터 머신 (Support Vector Machine; SVM)

SVM[5]은 각 클래스간 거리를 최대로 하는 경계선 또는 경계면(hyperplane)을 찾는다. 그리하여 새로운 데이터가 들어 왔을 때 일반화 오류를 최소화 하는 모델이다.

이때 각 클래스에서 데이터까지의 최소 거리를 마진(Margin), 그리고 경계선으로부터의 최소 거리인 데이터 벡터를 서포트 벡터(Support Vector)라고 한다. SVM은 해석이 용이하고, 성능이 뛰어나며, 적은 데이터에서도 적절한 결과가 나온다는 장점이 있다. 인공 신경망에도 크게 뒤지지 않는 성능을 내기에 오래도록 꾸준히 사용되고 있다.

2.3 컨볼루션 신경망 (Convolutional Neural Networks; CNN)

CNN[7]은 사람의 시신경망에서 아이디어를 얻어 고안한 모델로, 다양한 패턴 인식 문제에 사용되고 있다. CNN은 컨볼루션 층, subsampling 층 (또는 max-pooling 층)이라는 두 층을 번갈아가며 수행하다가 마지막에 있는 fully-connected 층을 이용하여 분류를 수행한다. 컨볼루션 층은 입력에 대해 2차원 필터링을 수행하고, subsampling 층은 매핑된 2차원 이미지에서 최댓값을 추출한다. 이러한 계층구조 속에서 역전파(backpropagation)를 이용, 오차를 최소화하는 방향으로 학습해나간다.

CNN은 패턴 또는 feature를 뽑는 문제에 뛰어난 성능을 보여 왔으며 주로 비전 분야에서 얼굴 인식[8], 필기체 인식[9] 등에 많이 사용되어 왔으나 최근에는 자연어 처리 분야에서도 널리 활용되고 있다.[10, 11]

2.4 순환 신경망 (Recurrent Neural Networks; RNN)

RNN[12]은 신경망 속 셀의 현재 출력 결과가 이전의 계산 결과에 영향을 받는 인공신경망 모델이다. 다시 말해, 이전 계산 결과에 대한 메모리 정보를 가지고 있어 순차적인 데이터를 학습하는데 장점을 가지고 있다. 기본적인 RNN은 일반적으로 학습이 어려워 다양한 변형이 발생했는데, 그 중 가장 성공적인 모델은 장기-단기 기억 신경망 (Long-Short Term Memory; LSTM)[13]과 최근 각광 받고 있는 회로형 순환 유닛 (Gated Recurrent Units; GRU)[14]이 있다.

LSTM은 긴 순차적인 정보를 회로 메커니즘 (gating mechanism)을 통해 저장하고 출력할 수 있다. 이 회로 메커니즘은 RNN의 학습을 방해하는 가장 큰 원인인 vanishing gradient 문제를 완화시켜 성능을 크게 향상시켰다.

2014년에 LSTM과 동일한 회로 메커니즘을 사용하지만 파라미터 수를 줄인 GRU가 제안되었다. GRU는 리셋 게이트와 업데이트 게이트로 구성되어 있으며, 두 게이트의 상호작용을 통해 학습한다. LSTM보다 적은 파라미터를 사용하기 때문에 이론적으로는 학습 속도가 조금 더 빠르고 완전한 학습에 필요한 데이터가 LSTM보다 적게 필요하다. 그러나 실제 성능으로는 특정 작업에서는 더 뛰어나기도 하고 뒤처지기도 한다[15].

3. 실 험

데이터 전처리와 Naive Bayesian Classifier, SVM은 Python2.7, CNN과 RNN은 Torch7의 nn, rnn 패키지[16]를 이용하여 구현하였다.

3.1 데이터

대분류로는 9개, 소분류로는 68개 분야에 분포되어 있는 인터넷에서 수집한 623,303개의 뉴스 데이터를 준비하였으며 학습 데이터, 검증 데이터, 테스트 데이터는 각각 70%, 15%, 15%의 비율로 나누었다. 데이터의 분야별 분포는 표 1과 같다.

3.2 설 계

비교모델로 TF-IDF (Term Frequency-Inverse Document Frequency)[17]를 사용한 Multinomial Naive Bayesian classifier와 SVM을 사용하였다.

CNN의 경우, 먼저 각 문서들을 형태소 분석기로 나눈 후 빈도수 기준 상위 n개의 단어로 lookup 테이블을 만

Large Category	Small Category(#Document)	합계
Science	Kistiscience(2,065), Science_general(3,063), Scienceskill(425)	5,553
Special Section	Esc_section(8,494)	8,494
International	Arabafrica(6,221), China(5,556), Globaleconomy(2,199), Europe(5,431), Internationalunit(564), Asiapacific(3,684), America(14,359), International_general(14,203), Globaltopic(887), Japan(6,437)	59,541
Economy	Working(2,852), Finance(6,228), Marketing(1,098), IT(3,998), Car(3,371), Stock(4,299), Heri_review(795), Biznews(6,393), Consumer(5,160), Property(5,970), Economy_general(49,679)	89,843
Politics	Bluehouse(5,856), Assembly(14,946), Politics_general(35,056), Defense(16,864) Administration(1,951), Diplomacy(2,926)	77,599
Culture	Travel(889), Movie(7,761), Book(14,068), Culture_general(12,067), Entertainment(13,054), Music(7,761), Religion(2,218), Skysea(50), Novel_salt(126), Baryprincess(126)	57,386
Society	Labor(6,305), Internalmove(459), Religious(2,448), Campus(4,338), Environment(5,249), Society_general(105,143), Women(1,604), Schooling(19,725), Health(7,484), Ngo(2,332), Rights(2,232), Obituary(4,398), Media(6,128), Area(52,459), Handicapped(797)	229,226
Opinion	Dica(1,331), Column(18,701), Because(5,462), Editorial(9,398), Argument(141)	35,033
Sports	Sports_general(24,113), Gameschedule(2,672), Baseball(11,911), Baduk(605), Scoreboard(382), Soccer(17,042), Golf(3,834)	60,559
총합 : [9 / 68 class] 623,303 documents		

표 1 실험 데이터 문서 수

든다. 그 다음 컨볼루션 커널을 슬라이드 하여 적절한 파라미터를 학습한다. 이때 커널의 가로와 세로 크기는 각각 단어 임베딩 크기, N-gram과 같이 동시에 학습하는 단어 크기와 같다. 활성 함수 (activation function)으로는 ReLU[18]를 사용했으며, logSoftMax를 이용하여 각 문서들이 특정 주제에 속할 확률을 출력하였다. 그 중 가장 높은 값을 가진 카테고리를 정답으로 예측하였다.

RNN의 경우, 마찬가지로 lookup 테이블을 생성하고, 테이블을 단어 벡터 단위로 쪼개어 순환 신경망의 입력으로 넣는다. lookup 테이블을 업데이트하면서 계속 해서 학습, 최종 은닉 층을 출력한다. CNN과 동일하게 ReLU와 logSoftMax를 이용하여 예측을 수행하였다.

3.3 실험 결과

각 모델 정확도는 표 2, 표 3과 같다.

Model	Accuracy (Top-1,3,5)		
MNB	0.641	0.911	0.958
SVM	0.795	0.960	0.991
CNN	**0.856**	**0.986**	**0.997**
LSTM	0.811	0.965	0.994
GRU	**0.886**	**0.992**	**0.999**

표 2 모델 별 대주제 실험 정확도

Model	Accuracy (Top-1,3,5)		
MNB	0.399	0.679	0.794
SVM	0.614	0.851	0.906
CNN	**0.700**	**0.920**	**0.962**
LSTM	0.670	0.895	0.942
GRU	**0.725**	**0.937**	**0.971**

표 3 모델 별 소주제 실험 정확도

실험 결과, GRU가 가장 뛰어난 성능을 보였다. LSTM은 SVM과 CNN 사이의 성능을 보였다.

4. 결 론

본 논문은 인터넷에서 수집한 텍스트 문서를 여러 알고리즘을 이용하여 정해진 카테고리에 맞게 분류하는 내용을 담고 있다. 총 623,303개의 문서를 대분류 9개, 소분류 68개에 분류하였을 때, 분류 정확도는 Multinomial Naive Bayesian Classifier, SVM, LSTM, CNN, GRU의 순서로 나타났다. 이 결과는 다음과 같이 해석할 수 있다: (1) LSTM보다 CNN의 성능이 더 뛰어난 것으로 보아 문서 분류 문제는 전체 글의 시퀀스를 학습하는 것 보다는 글의 feature를 통해 학습하는 것이 더 올바른 문제 접근법이라고 생각할 수 있다. 따라서, (2) LSTM과 GRU의 성능을 비교했을 때, LSTM에 비해 GRU가 feature를 더 잘 추출했다고 볼 수 있다. 마지막으로, (3) GRU가 CNN보다 성능이 더 좋은 것으로 보아, 문서 분류 문제는 feature와 시퀀스 두 가지를 모두 적절히 고려할 때 성능이 가장 잘 나온다는 것을 확인할 수 있었다. 추후 연구에서는 LSTM과 GRU에서 추출된 feature와 embedding 결과를 비교, 분석하여 어떤 차이가 GRU의 성능을 더 뛰어나게 만들었는지 확인해보고자 한다.

감사의 글

이 논문은 2015년, SK텔레콤의 지원을 받아 수행된 연구임.

참고 문헌

[1] Chai, Kian Ming Adam, Hai Leong Chieu, and Hwee Tou Ng. "Bayesian online classifiers for text classification and filtering." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.

[2] Jovanovic, Jelena, et al. "Automated Semantic Tagging of Textual Content." IT Professional 16.6 pp.38-46, 2014.

[3] Harris, Zellig S. "Distributional structure." Word 10.2-3, pp.146-162, 1954.

[4] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." Machine learning 29.2-3, pp.131-163, 1997.

[5] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992.

[6] Jo, Hwiyeol, Kim, Jin-Hwa, Yoon, Sangwoong, Kim, Kyung-Min and Zhang, Byoung-Tak, Large-Scale Text Classification with a Convolutional Neural Network. 42th The Korean Institute of Information Scientists and Engineers Annual Meeting, 2015.

[7] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series," In M. A. Arbib (Ed.), The handbook of brain theory and neural networks, Cambridge, MA: MIT Press, pp. 255-258, 1995

[8] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[9] Bluche, Théodore, Hermann Ney, and Christopher Kermorvant. "Feature extraction with convolutional neural networks for handwritten word recognition." Document Analysis and Recognition (ICDAR), 12th International Conference on. IEEE, 2013.

[10] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." arXiv preprint arXiv:1404.2188, 2014.

[11] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification", Empirical Methods on Natural Language Processing, 2014

[12] Goller, Christoph, and Andreas Kuchler. "Learning task-dependent distributed representations by backpropagation through structure." Neural Networks, IEEE International Conference on. Vol. 1. IEEE, 1996.

[13] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8, pp.1735-1780, 1997

[14] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014.

[15] Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. "An empirical exploration of recurrent network architectures." Proceedings of the 32nd International Conference on Machine Learning (ICML-15). 2015.

[16] Léonard, Nicholas, Sagar Waghmare, Yang Wang, and Jin-Hwa Kim. rnn: Recurrent Library for Torch. arXiv preprint arXiv:1511.07889, 2015.

[17] Yun-tao, Zhang, Gong Ling, and Wang Yong-cheng. "An improved TF-IDF approach for text classification." Journal of Zhejiang University Science A 6.1, 2005.

[18] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010.