문장 유사도 측정을 위한 Siamese Network 연구

이민석, 이창우
국민대학교 소프트웨어학부
e-mail : lms990427@kookmin.ac.kr

Siamese Network for Sentence Similarity Measurement

Lee Min Seok, Lee Chang Woo
Dept of Computer Software, Kookmin University

요 약

본 논문은 Siamese CNN 신경망을 이용한 문장추출기를 만들어 두 문장간의 유사도를 추출할 수 있도록 하는 방법을 제시한다. 기존 연구에서는 Siamese Network 가 이미지 유사도에 많이 활용되었다. 하지만 텍스트 분석은 이미지 분석과는 다른 특성을 가지며, 문장 구조, 단어의 의미, 문맥 등을 고려해야 한다. 본 연구에서는 두 데이터의 차이점을 분석하는데 좋은 성능을 보이는 Siamese Network 를 이미지가 아닌 텍스트 분석에 사용하여 연구한다.

1.서론

지금까지 이미지의 유사도 추출에는 두 네트워크의 파라미터를 공유하여 유사도를 구하기에 적합한 Siamese Network 를 많이 이용해 왔다[1]. 이는 얼굴인식, 이상징후 탐지 등 여러 분야에 사용되어왔다. 하지만 최근 연구는 텍스트에 적용하려는 사례가 있다[2].

이 연구의 주요 목표는 Siamese CNN 신경망을 활용하여 두 문장 간의 유사도를 추출하는 방법을 제시하고, 이미지 분야에서의 Siamese Network 와 비교하여 성능을 평가하는 것이다. 본 논문은 서론, Siamese Network 의 구조, 실험 및 결과, 결론 등으로 구성한다.

2.본론

2.1. Siamese Network 의 기본 구조

Siamese Network 는 그림 1 처럼 두개의 입력값으로 이루어진다. 이 두개의 입력을 받기 위해서는 두개의 동일한 서브 네트워크로 구성하고, 이 네트워크들은 동일한 파라미터를 공유하며, 두개의 다른 입력 문장을 처리한다. 또한 이 네트워크의 출력은 두 입력의 유사성을 측정하는 거리 메트릭 에 기반하여 나오게 되며 이 거리가 작을 수록 두 문장은 더 유사하다고 판단하게 된다. 따라서 유사한 두 문장이 입력값으로 들어갈때는 출력값이 최대값이 되도록, 반대의 경우에는 최소값이 되도록 학습을 반복적으로 진행한다[3].

2.2 네트워크 아키텍처

CNN(Convolution Neural Network) 계층은 주어진 입력 텐서의 차원이 (배치 크기 x 1 x 시퀀스 길이 x 임베딩 차원)일 때, 아키텍처는 여러 필터 크기로 여러 Convolution Layer 를 적용한다. 특히, 필터의 차원은 (필터 크기, 임베딩 차원)이다.

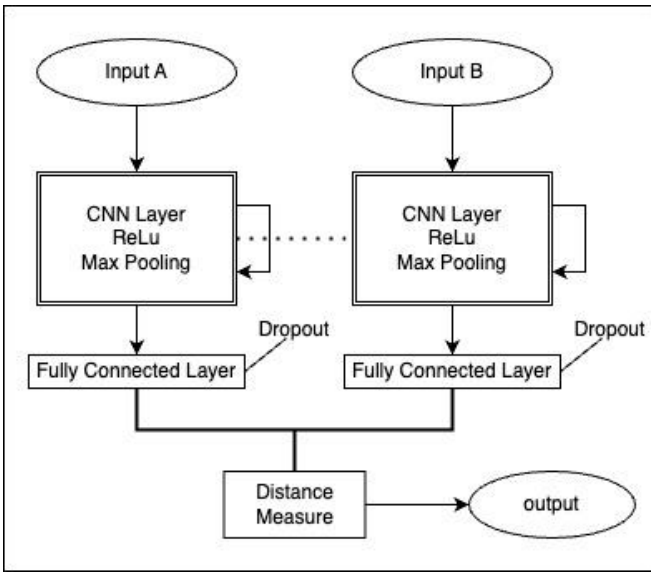

그림 1. Siamese Network

컨볼루션 연산 후에는 ReLU 활성화 함수가 적용된다. 맥스 풀링은 컨볼루션 레이어 다음으로, 네트워크는 출력에 최대 풀링 연산을 적용한다. 최대 풀링의 차원은 (64, 1)이다. 맥스 풀링후 각 컨볼루션 레이어의 출력은 연결되고 완전 연결 레이어를 통과한다. 이 레이어 후에는 ReLU 활성화 함수가 적용된다.

이후에는 과적합을 피하기 위해 0.5 의 드롭아웃 비율로 드롭아웃 레이어가 적용된다. 드롭아웃 레이어의 출력은 최종 임베딩을 생성하기 위해 다른 완전 연결 레이어를 통과하여 출력 레이어를 생성한다.

2.3 훈련

Siamese Network 는 두개의 데이터가 동시에 들어가야하므로 데이터를 쌍을 지어서 넣어주어야 한다. 또한 학습을 위한 데이터는 해당 데이터의

레이블을 넣어주었다. 본 논문에서 사용한 데이터는 직접 크롤링한 리뷰데이터를 이용했다. 각각 1~5 점의 별점을 가지는 리뷰 1000 개를 이용하였다. 데이터의 임베딩은 한국어 처리에서 가장 빠른 속도와 성능이 높은 GloVe 방식을 이용하였다[4].

GloVe 방식은 대규모 텍스트 코퍼스 내의 단어 공동 출현 통계를 활용하여 단어들의 의미를 밀집 벡터로 표현한다. GloVe 의 핵심 아이디어는 단어 쌍의 공동 출현 확률과 그들의 벡터 내적 사이의 관계를 최적화하는 것이다. 이를 통해, 단어 간의 의미적 관계와 문법적 속성을 더 잘 포착하는 임베딩이 생성될 수 있다. GloVe 는 기존의 카운트 기반 방식과 예측 기반 방식의 장점들을 모두 결합하여, 효과적으로 단어의 의미를 벡터로 임베딩 할 수 있다[5].

같은 별점대의 리뷰는 유사하다 라는 가정을 하고 리뷰 데이터는 학습데이터와 테스트데이터로 나누어 각각 학습과 테스트에 사용하고, 댓글은 해당 데이터의 레이블로 사용하였다.

Siamese Network 를 학습할때, 하이퍼파라미터로 임베딩차원, 필터의 개수, 필터의 크기, 은닉차원, 학습률, 배치크기, 반복횟수(epochs)를 설정한다.

Siamese Network 를 학습시키는 과정은 모델을 학습 모드로 설정하고 옵티마이저의 그래디언트를 초기화 한다. 학습 데이터의 두 데이터와, 이와 대응하는 레이블 을 가져와 모델에 전달한다. 모델에서 두 입력에 대한 출력을 계산하고, Margin Ranking Loss 를 이용하여 손실을 계산한다. 손실에 역전파를 수행하고, 옵티마이저를 사용하여 모델의 가중치를 업데이트한다.

2.4 결과 도출

Siamese 네트워크를 학습하고 나면, 테스트 데이터셋을 사용하여 모델의 성능을 평가했다. 학습된 모델을 사용하여 두 입력 문장 사이의 유사도를 계산하고, 이를 바탕으로 이진 예측을 생성했다. 여기서 예측 임계값은 0.5 로 설정했다.

테스트셋에 대한 평가 결과 Siamese Network 의 정확도는 54%를 달성했다. 그림 2 에서 Loss 값이 일정 범위 밑으로는 내려가지 못하는 모습을 보인다. 데이터에 차이는 있겠지만 30k 의 훈련을 거친 이미지 분류기가 약 90% 이상의 정확도를 보이는것에 비하면 매우 낮은 수치로 판단한다[3].

수치가 낮은 원인은 Siamese Network 는 일반적으로 데이터의 지역적, 전역적인 패턴을 포착하여 유사도를 검사하는 만큼, 하나의 패턴이 하나의 의미를 가지는 이미지 보다 하나의 패턴이 여러 의미를 가지는 텍스트에 대한 분석이 더 정확하지 않은 결과가 나온 것으로 예상된다.

3.결론

본 논문에서 진행한 연구에서는 Siamese Network 를 활용하여 텍스트 유사도 측정 모델을 실험하고, 텍스트 간의 유사성 평가를 수행하였다. 실험 결과로부터 Siamese Network 를 통해 텍스트 간의 유사성을 일정 수준으로 평가할 수 있음을 확인한다 그러나 현재의 성능은 이미지 분류와 비교하여 낮은 수준임을 인지하고 있다.

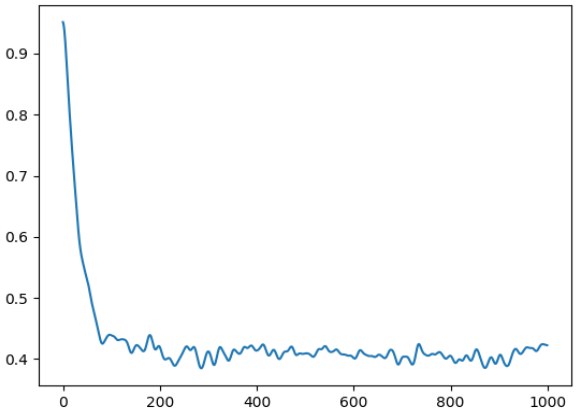

그림 2. 테스트에서의 Loss 값 변화

향후에 각 레이블별로 더 일관적인 특성을 가지는 많은양의 학습 데이터를 통한 학습의 정확성 향상과, 임베딩 단계에서의 전처리를 통한 데이터의 특성 파악 용이하게 하고,또 다른 추가적인 파라미터의 조정을 통해 퇴적의 성능을 추구하여 더 좋은 결과를 낼 수 있도록 연구하고자 한다.

ACKNOWLEDGEMENT

본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음. (2022-0-00964)

참고문헌

[1] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, ‘‘Signature verification using a ‘siamese’ time delay neural network,’’ in Proc. Adv. Neural Inf. Process. Syst., Feb. 1993, pp. 737-744.

[2] L. Yan, Y. Zheng, and J. Cao, “Few-shot learning for short text classification,” Multimedia Tools Appl., vol. 77, no. 22, pp. 29799-29810, 2018.

[3] G. Koch, “Siamese neural networks for one-shot image recognition,” M.S. thesis, Dept. Comput. Sci., Toronto Univ., Toronto, ON, Canada, 2015.

[4] Juree Seok, Heuiseok Lim, “Word2Vec, GloVe 및 RoBERTa 등의 모델을 활용한 한국어 문장 임베딩 성능 비교 연구,” 제 33 회 한글 및 한국어 정보처리 학술대회 논문집, pp. 444-449, 2021.

[5] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), 2014, pp. 1532-1543.