문장 유사도 검사를 위한 Siamese Network 연구

이민석, 이창우
국민대학교 소프트웨어학부

서론

◆ ***Siamese CNN* 신경망의 새로운 적용**: 이미지 분석에 주로 사용되는 Siamese CNN 신경망을 텍스트 분석에 적용해 문장 간 유사도를 측정하는 방법을 탐구

◆ **텍스트의 복잡성 고려**: *Siamese Network*가 문장 구조, 단어 의미, 문맥 등 텍스트의 복잡한 요소들을 고려하여 텍스트 분석에 어떻게 적용될 수 있는지 연구

◆ **텍스트 분석에서의 도전**: *Siamese Network*를 텍스트에 적용하는 것은 이미지 분석과는 다른 도전을 제시하며, 이 연구는 이러한 도전에 대응하는 방법을 모색

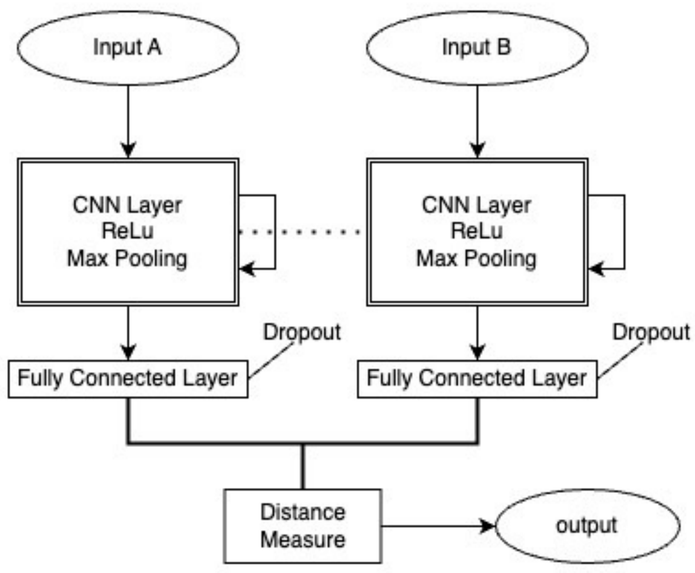
◆ **성능 평가**: 연구의 핵심 목표는 문장 간 유사도를 추출하는 *Siamese CNN* 신경망의 성능을 평가하고, 이를 이미지 분석에 사용된 사례와 비교

◆ **연구 구성**: 해당 연구는 서론, *Siamese Network*의 구조 설명, 실험 및 결과, 결론 등으로 구성

◆ ***Siamese Network*의 텍스트 분석 적용 가능성**: 이 연구는 *Siamese Network*가 텍스트 분석, 특히 텍스트 유사도 측정 분야에서 새로운 가능성을 제시

◆ **텍스트 데이터 이해 증진**: *Siamese Network*가 텍스트 데이터 내의 다양한 패턴과 의미를 어떻게 처리하고 이해할 수 있는지에 대한 이해를 높이는 것을 기대

본론



Siamese Network의 기본 구조

◆ Siamese Network는 두 개의 서로 다른 입력을 받아 각각의 동일한 서브네트워크에서 처리, 이들의 파라미터를 공유하여 문장의 유사도를 판단.

◆ 네트워크는 출력 거리 메트릭을 통해 두 문장의 유사성을 측정, 출력 거리가 짧을수록 문장이 유사하다고 판단.

◆ 학습 과정에서 네트워크는 유사한 문장을 더 잘 구별할 수 있도록 출력 거리의 차이를 극대화하는 방향으로 조정.

네트워크 아키텍쳐

1. CNN 계층 구성: CNN 계층은 여러 크기의 필터를 사용하는 다중 컨볼루션 레이어로 구성됨. 각 필터는 입력 텐서의 임베딩 차원과 매칭.

2. 활성화 함수 적용: 컨볼루션 연산 이후 각 레이어에는 ReLU 활성화 함수가 적용.

3. 풀링과 드롭아웃: 최대 풀링 연산은 차원을 (64, 1)로 줄인다, 과적합 방지를 위해 드롭아웃 레이어가 적용. 드롭아웃 비율은 0.5이다.

4. 네트워크의 최종 단계: 컨볼루션 레이어의 출력은 완전 연결 레이어를 거쳐 드롭아웃을 적용한 후 최종 출력 레이어로 전달.

훈련

1. 데이터 준비: 훈련을 위해 1에서 5점까지 별점이 있는 1000개의 한국어 리뷰 데이터를 쌍으로 크롤링하여 준비.

2. 임베딩 방식: 리뷰 데이터는 단어의 공동 출현 통계를 기반으로 하는 GloVe 방식을 사용하여 의미적, 문법적 속성을 포착하는 밀집 벡터로 임베딩.

3. 데이터 분할 및 레이블링: 별점이 같은 리뷰들을 유사한 것으로 간주하며, 데이터를 학습과 테스트 세트로 나누고, 별점은 학습 데이터의 레이블로 사용.

4. 하이퍼파라미터 설정: 학습 과정에서 임베딩 차원, 필터 수, 필터 크기, 은닉 차원, 학습률, 배치 크기, 반복 횟수 등 다양한 하이퍼파라미터를 설정.

5. 훈련 과정: *Siamese Network*의 훈련은 학습 모드 설정, 옵티마이저 초기화, 학습 데이터와 레이블 전달, 출력 계산, *Margin Ranking Loss*를 사용한 손실 계산 및 역전파를 통한 가중치 업데이트로 구성.

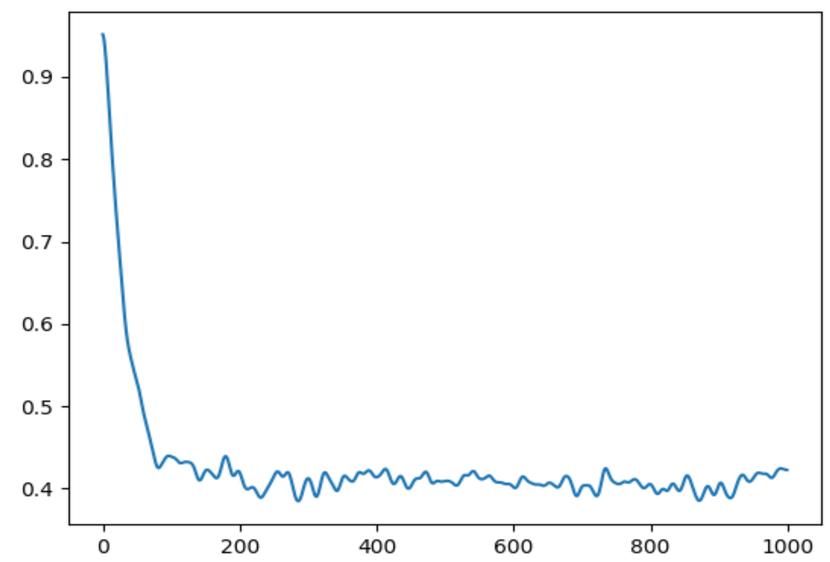
결과도출

◆ **테스트 결과**: *Siamese Network*는 테스트 데이터셋에서 약 54%의 정확도를 나타내어, 90% 이상의 정확도를 달성하는 훈련된 이미지 분류기와 비교했을 때 상대적으로 낮은 성능을 보임.

◆ **성능 한계**: 예측 임계값을 0.5로 설정했음에도 불구하고, Loss 값이 일정 수준 이하로 떨어지지 않는 경향을 보여, *Siamese Network*가 텍스트 데이터 처리에서 한계를 보이는 것으로 나타남.

◆ **유사도 검사의 복잡성**: 텍스트 데이터에서 *Siamese Network*의 낮은 정확도는 텍스트에서 단일 패턴이 다양한 의미를 가질 수 있는 복잡성 때문에 발생했을 것으로 추측.

결론



X : epoch 수 , Y : Loss값

텍스트 유사도 측정을 위해 Siamese Network를 사용한 모델을 실험하고 평가했다. 실험 결과, Siamese Network는 텍스트 간의 유사도를 어느 정도 정확하게 평가할 수 있음을 보여줬지만, 이미지 분류에 비해 낮은 성능을 나타냈다.

이에따라 레이블별로 일관된 특성을 가진 더 많은 학습 데이터를 활용하여 모델의 정확도를 높이려고 한다. 또한, 임베딩 과정에서 데이터 전처리를 통해 특성을 더 잘 파악하고, 추가적인 하이퍼파라미터 조정을 통해 성능을 개선하고자 한다. 이를 통해 텍스트 분석 분야에서 Siamese Network의 더 나은 결과를 도출하기 위한 연구를 계속 진행할 계획이다.