자연어 처리(NLP)에서 지도 학습에 대한 의존도를 줄이기 위해, 원시 텍스트로부터 효과적으로 학습하는 능력이 매우 중요하다. 대부분의 딥러닝 방법은 많은 양의 수작업으로 라벨링된 데이터가 필요하며, 이는 주석 자원이 부족한 많은 분야에서 적용을 어렵게 만든다. 이런 상황에서는 라벨이 없는 데이터로부터 언어 정보를 활용할 수 있는 모델이 시간과 비용이 많이 드는 수작업 주석 작업을 대신할 수 있다.

뿐만 아니라, 충분한 감독 학습 데이터가 있는 경우에도 비지도 방식으로 좋은 표현을 학습하는 것은 성능 향상에 큰 도움이 된다. 이의 대표적인 사례가 사전 훈련된 word embedding의 광범위한 사용으로, 다양한 NLP 작업의 성능을 개선해왔다.

그러나 단어 수준을 넘는 정보들을 라벨 없는 텍스트로부터 학습하는 것은 어렵다. 학습하기에 어려운 이유는 총 2가지가 있다.

첫 번째 이유는 어떤 학습 목표objective가 전이 학습에 유용한 표현을 학습하는 데 가장 효과적인지 확실히 알 수가 없기 때문이다. 최근의 연구들은 언어 모델링, 기계 번역, 담화 일관성 등 다양한 목표를 시도했고, 각각 다른 작업에서 서로를 능가하였다.

두 번째로는, 학습된 표현을 실제 작업에 어떻게 전이할 것인지에 대해서도 확실하지 않기 때문이다. 현재의 기법들은 작업별로 모델 구조를 변경하거나, 복잡한 학습 방식, 추가적인 보조 학습 목표 등을 포함하고 있다.

이러한 불확실성은 효과적인 반지도 학습 접근법 개발을 어렵게 한다.

본 논문은 비지도 사전 학습과 지도 미세 조정을 결합한 반지도 방식으로 언어 이해 작업에 대해 다룬다. 논문의 목표는 적은 수정만으로도 다양한 작업들에 전이가 가능한 universal representation을 학습하는 것이다.

1. 대규모의 라벨 없는 텍스트와 여러 개의 수작업 라벨이 있는 데이터셋(타겟 작업)에 접근할 수 있다고 가정한다.
2. 이때, 타겟 작업은 라벨 없는 텍스트와 같은 도메인일 필요는 없다.
3. 학습은 사전 학습 단계와 미세 조정 단계로 구성이 된다.

   먼저, 사전 학습의 경우 라벨 없는 데이터에 대해 언어 모델링 목표를 사용하여 신경망의 초기 가중치를 학습하는 과정이고, 미세 조정 단계의 경우 각 타겟 작업에 맞는 감독 학습 목표로 모델을 조정하는 과정이다.

모델로는 Transformer 아키텍처를 사용한다. Transformer 모델은 기계 번역, 문서 생성, 구문 분석 등 다양한 작업에서 강력한 성능을 보인 모델이다. 순환 신경망보다 긴 거리의 의존 관계를 더 잘 처리할 수 있어, 다양한 작업에서 더 안정적인 전이 성능을 제공하고, 전이 과정에서는 structured text를 하나의 연속된 토큰 시퀀스로 처리하는 traversal-style 입력 방식을 사용하여 사전 학습된 모델 구조를 거의 바꾸지 않고도 효과적으로 잘 미세 조정할 수 있다.

자연어 추론, 질문 응답, 의미 유사도, 텍스트 분류 이렇게 4가지 분야에서 평가를 진행하였다. 평가 결과, 일반적인 모델임에도 불구하고, 각 작업에 특화된 모델들을 뛰어넘는 성과를 냈다는 사실을 알 수 있었다. Commonsense reasoning의 경우 전보다 8.9%, Question answering의 경우 전보다 5.7% 등 다양한 좋은 성과를 냈다. 또한, zero-shot 전이 실험에서도 유의미한 언어적 지식을 습득했다는 결과를 얻을 수 있었다.