0. Abstract

본 논문에서는 대규모 라벨 없는 텍스트로 비지도 사전학습을 수행한 뒤, 소량의 라벨 데이터로 지도 미세조정을 진행하는 Generative Pre-Training(GPT-1) 방식을 제안하였다. 이 접근법을 통해 범용적이며 다양한 자연어처리 태스크에서 우수한 성능을 달성하였다.

1. Introduction

딥러닝 기반 NLP는 대부분 레이블이 부착된 데이터에 의존해 학습한다. 그러나 레이블링 비용이 크기 때문에 이를 줄이고, 라벨이 없는 대량의 텍스트에서 유용한 표현(언어적 정보)을 효과적으로 학습하는 비지도/반지도 전략의 중요성이 강조되고 있다. 특히, 다양한 태스크에서 잘 작동하는 최적화 목표와 효과적인 전이학습 방식이 중요한 연구 이슈였다.

2. Background: Generative/Discriminative/Transfer Learning

- Generative Learning: 레이블의 분포까지 포함해 데이터 전체 분포를 학습하는 방법이다. 데이터가 많을수록 모델링 결과가 풍부해진다.
- Discriminative Learning: 레이블에 따라 경계(Decision boundary)를 학습한다. 데이터가 적어도 동작하지만 과적합 위험이 크다.
- Transfer Learning: 비지도 임베딩(Word2Vec, ELMo 등)을 활용하거나, 소량의 라벨 데이터와 함께 쓰는 반지도학습 등이 적극적으로 연구되고 있었다.

3. Main Idea: Generative Pre-trained Transformer

GPT는 다음과 같은 2단계로 학습이 이뤄진다.

1. Unsupervised Pre-training
    - BooksCorpus(책 7,000권 이상), 1B Word Benchmark 같은 대규모 라벨 없는 텍스트 코퍼스에 대해 언어모델링(다음 단어 예측) 방식으로 transformer(디코더 구조)를 사전학습시켰다.
    - 이 과정에서 긴 문맥 정보를 효과적으로 포착하기 위해 word embedding, positional encoding, BPE(Byte-Pair-Encoding)가 사용됐다.
    - 언어모델의 목표함수는 log likelihood를 최대화하는 방식이며, context window(문맥 토큰 수)의 크기도 실험적으로 다뤘다.
2. Supervised Fine-tuning
    - 사전학습된 모델 파라미터를 활용하여, 각 작업마다 task-specific dataset(분류, NLI, QnA 등)에 맞게 소량의 라벨 데이터로 미세조정하였다.
    - 입력 데이터의 포맷만 태스크에 맞게 변환(예: 분류, entailment, 유사도, 다

중선택 등 각각 형식에 맞는 구성)해서 구조 변경 없이 여러 작업을 하나의 모델로 해결하도록 했다.

- 미세조정 단계에서는 pre-training에서 사용했던 보조 목적함수(auxiliary training objective)를 같이 써서 일반화 성능을 높이고 학습 속도도 개선하였다.

4. Task-specific Input Transformation

GPT의 입력 포맷 변환(Task-specific Input Transformation)은 다음과 같다.

- 분류(Classification): 감정분석, 뉴스, 문서분류 등에서 입력 텍스트를 transformer에 넣고 hidden state를 선형결합한다.
- 의미적 포함(Entailment, NLI): premise(전제)와 hypothesis(가설) 두 문장을 delimiter로 구분해 입력한다.
- 유사도(Similarity): 두 문장 순서를 바꿔 각각 transformer에 입력한 결과를 concat하여 최종 유사도를 산출한다.
- 다지선다(Multiple Choice): 질문과 여러 후보 정답을 각각 별도 입력으로 넣어 softmax 확률 분포로 최종 정답을 고른다.

이처럼 다양한 태스크에 맞게 입력만 다르게 하며, 모델 구조의 변화 없이 범용적으로 활용했다.

5. 실험 및 결과(Experiment & Results)

- 사전학습에 BooksCorpus, 1B Word Benchmark를 사용했다. BooksCorpus는 긴 문맥 구조를 고려한 코퍼스라서 장기 정보 처리에 유리했다.
- BPE(Byte-Pair-Encoding) tokenizer를 사용해 신조어, 복합어, 파생어 등 실제 언어 현상을 잘 포착했다.
- GLUE, MNLI, SNLI, QNLI 등 여러 벤치마크(task)에서 SOTA에 가까운 성능을 기록하였다.
- 자연어추론, 분류, 질의응답 등 다양한 작업마다 별도 구조 변경 없이 동일 모델·파라미터를 사용했고, 실험에서 앙상블 SOTA 모델보다도 우수한 결과를 얻었다.

6. Ablation Study & 분석

- pre-training 없이 supervised만 사용한 경우, pre-training을 거친 경우에 비해 성능이 일관되게 떨어지는 것을 확인했다.
- LSTM 계열 구조보다 transformer(decoder-only)의 장기 정보 파악력이 뛰어나다는

점을 실증했다.

- 사전학습 후 얼마나 많은 레이어(Depth)를 전이할수록 성능이 오르는지도 실험을 통해 밝혔다.
- fine-tuning 시 pre-train에서 얻은 파라미터들을 고정하지 않고, 계속 갱신하며 목적함수의 가중치(람다)를 조정해 task overfitting을 억제했다. 논문은 람다 0.5를 사용했다.

7. Zero-shot Behavior

- GPT는 한 번도 훈련하지 않은 태스크에서도 적절한 포맷으로 입력을 주면(Zero-shot), 일정 수준의 성능을 내며, pretrained transformer가 LSTM보다 다양한 task에서 탁월했다.
- pre-training 횟수를 늘릴수록 전반적인 generalization 성능이 높아짐을 보였다.

8. 결론(Conclusion)

GPT-1은 대규모 비지도 사전학습 + 소량 내지는 태스크별 지도 미세조정을 통해, 입력 포맷 변환만으로 다양한 자연어처리 task를 범용적으로 처리하는 패러다임을 처음 제시했다. 이는 이후 BERT, GPT-2, 대형 LLM 연구에 큰 영향을 주었다