Week 3

Attention Is All You Need (NLP)

https://arxiv.org/pdf/1706.03762

Abstract

이 문서는 'Transformer' 모델에 대한 자세한 설명을 제공합니다. Transformer는 전통적인 순환 신경망(RNN)이나 합성곱 신경망(CNN)을 사용하지 않고, 전적으로 주의(attention) 메커니즘에 기반한 새로운 네트워크 아키텍처로, 더 나은 품질과 빠른 학습속도를 자랑합니다. 이 연구는 기계 번역 작업에서 기존의 모델을 초월하는 성능을 입증했습니다. 이 논문을 통해 독자는 최신 자연어 처리 기술의 기초를 이해하고, Transformer가 어떻게 글로벌 종속성을 모델링하며 기존 접근 방식을 개선하는지를 배울 수 있습니다.

핵심 Contents

1. **도입 및 동기 (2017년 초)**:
    - **배경**: 이전의 시퀀스 변환 모델들은 주로 순환 신경망(RNN) 또는 합성곱 신경망(CNN)에 기반했으며, 이는 계산 비용이 높고 병렬 처리가 어렵다는 한계가 있었습니다.

- **핵심 아이디어**: 저자들은 **Transformer**라는 새로운 모델을 소개했습니다. 이 모델은 시퀀스 내 요소들 간의 의존성을 처리하기 위해 **자기-어텐션(self-attention)** 메커니즘만을 사용하며, 순환이나 합성곱을 완전히 배제했습니다.

2. **모델 개발 및 구조 (2017년 중반)**:
   - **구조**: Transformer 모델은 인코더와 디코더 스택으로 구성되며, 각 스택은 여러 층의 자기-어텐션과 포인트-와이즈 완전 연결 레이어로 이루어져 있습니다. 인코더는 입력 시퀀스를 연속적인 표현으로 변환하고, 디코더는 이를 바탕으로 출력 시퀀스를 생성합니다.
   - **어텐션 메커니즘**: Transformer는 **Scaled Dot-Product Attention**과 **Multi-Head Attention**을 도입하여 다양한 시퀀스 위치 간 의존성을 효율적으로 학습하고 처리할 수 있습니다.
3. **훈련 및 성능 (2017년 후반)**:
   - **훈련 효율성**: Transformer는 RNN 또는 CNN 기반 모델보다 훈련 속도가 빠르며, 대규모 데이터셋에서 탁월한 성능을 보였습니다. 예를 들어, **영어-독일어 번역** 작업에서 기존 최고 성능 모델보다 더 높은 BLEU 점수를 기록했습니다.
   - **최적화**: 저자들은 Adam 옵티마이저와 학습률 조정 기법을 사용하여 성능을 극대화하고, 드롭아웃과 라벨 스무딩 등의 정규화 방법을 적용하여 과적합을 방지했습니다.
4. **결과와 응용 (2017년 말)**:
   - **번역 작업**: Transformer는 **WMT 2014 영어-독일어** 및 **영어-프랑스어** 번역 작업에서 새로운 최고 BLEU 점수를 기록하며, 이전 모델들보다 우수한 번역 품질을 보여주었습니다.
   - **다른 작업으로의 일반화**: Transformer는 영어 문장 구조 분석과 같은 다른 작업에서도 높은 성능을 보여, 번역 외에도 다양한 시퀀스 처리 작업에 적용 가능함을 입증했습니다.

Conclusion

저자들은 Transformer가 주목할 만한 병렬 처리 성능과 훈련 효율성을 통해 기존의 RNN 및 CNN 기반 모델들을 대체할 수 있는 잠재력을 갖추고 있다고 결론지었습니다. 향후 Transformer를 텍스트 이외의 입력과 출력, 예를 들어 이미지나 오디오 처리에 적용하는 연구 방향도 제시했습니다.