Long-Span Summarization via Local Attention and Content Selection

Abstract

긴 문서 요약에서 large scale의 pre-trained model을 적용할 때 발생하는 memory와 time 문제 발생 -> local self-attention, content selection 방법 사용한 요약 수행

1. Introduction

Transformer-based models: 많은 양의 데이터와 최적화 수행 -> issue: 긴 문서 요약에서 sequence length가 quadratically하게 자라는 문제

Quadratic한 특성을 해결하기 위한 modified self-attention mechanism과 transformer의 개선 -> 모델 가중치의 낮은 접근가능성 -> BERT나 BART같은 standard models는 다양한 task에 적용 가능하고, 적은 학습 시간 대비 훌륭한 성능을 보이므로 긴 문서 요약 task로 채택.

긴 문단의 의존성을 다루기 위한 두 가지 방법

Local self-attention: Input 길이를 늘리기 위해 attention mechanism을 local 단위로 제한하는 standard transformer models

Content selection: 문장에 순위를 매기기 위한 Multitask content selection(MCS) 방법 사용

arXiv, PubMed datasets으로 실험 착수, ROUGE scores로 평가, small-scale GPU card 사용

2. Related Work

Efficient Transformers

Pre-trained transformer models

- BERT in contextual representation
- GPT2 in text generation
- BART in seq2seq tasks

➔ Sequence length가 기하학적으로 늘어나는 긴 문서에 대해서 메모리 문제 발생

Full self-attention mechanism

- Fixed attention patterns

- Learnable patterns
- Low-rank matrix approximation
- Kernel method

불필요한 attention heads가 정리되어야 함

Knowledge distillation은 메모리와 연산량을 줄일 수 있다.

Encoder-decoder architectures

➔ Long input의 의존성(긴밀성)과 target sequences에 초점

Long-span Summarization

- BigBird
- Longformer-Encoder-Decoder(LED)
- Hierarchical transformer architectures (multi-document summarization)
- Extractive news and table-to-text summarization
- Hierarchical attention RNN system (summarize long articles)
- Content selection (news summarization systems)
- Extractive system + TLM (for scientific articles)
- Simple selection + BART (for podcasts)
- BERT-based keyword/sentence extraction + BART (for news and scientific articles)
- Dividing the source and target into multiple smaller pairs (to train abstractive summarizers)
- Extractive methods with and without redundancy re-duction techniques

3. Experimental Setup

A. Dataset

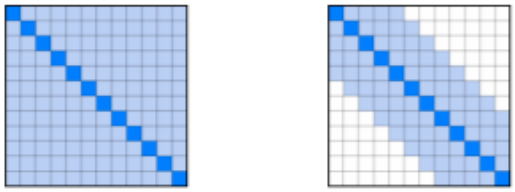
Spotify Podcast

arXiv and PubMed

B. Models

BART and LoBART

Encoder의 self-attention mechanism을 local self-attention BART로 적용.



(a) Full (b) Local (W=9)

Figure 2: Self-Attention Pattern.

Hierarchical RNN

기존의 content selection model

➔ Hierarchical encoder-decoder 기반

➔ Word-level, sentence-level GRUs로 구성된 모델

Label 추출을 위해 sentence-level GRU에 linear layer 추가

➔ Multitask content selection (MCS)

4. Longer Span via Local Self-Attention

A. Memory Analysis and LoBART Design

B. BART and LoBART

5. Longer Span via Content Selection

A. Training-time Content Selection

B. Multitask Content Selection(MCS)

6. Combined Approach

A. Spotify Podcast results

B. arXiv and PubMed results

C. Local Attention v.s. MCS

7. Conclusion

Reference