NLP Homework 2

영어 문서 요약

영어 문서 요약

- Summarization dataset

  - 영어 문서 요약을 위한 데이터셋

  - 1개의 document와 해당 document를 잘 요약하는 summary로 구성되어 있음.

  - 본 과제에서는 (document, summary) 쌍으로 이루어진 Train/Valid/Test 데이터셋과 document만 존재하는 Submit 데이터셋을 제공.

  - 예시

(Document) Because sun's rays are diffuse, solar panels must occupy substantial territory to generate any significant quantities of power. As a result, solar energy is land-intensive and creates a pressure to clear land of trees and vegetation to make way for solar panels. Owners of solar panels on home rooftops may also have an incentive to cut-down trees that are blocking solar panels from the sun's rays. This is a significant ecological threat.
(Gold Summary) Land-intensive solar power incentivizes clearing land ecosystems.

Introduction

- Homework 2

  - Transformer 모델을 사용해 document summarization 모델 학습 및 평가

    1. Summarization task에 맞게 데이터 전처리
    2. Transformer 모델 구현
    3. 요약 모델 학습 및 평가

1. Preprocessing (datamodule.py)

- Tokenizer를 이용하여 데이터 전처리

  - _convert_to_feature

    : 구현된 Tokenizer를 바탕으로, 데이터셋의 document, summary가 입력되면 이를 tokenizing -> id로 변환 -> 이후의 전처리 (truncation / padding) 해주는 method.

    딥러닝 모델의 원활한 학습을 위해서 입력되는 token들의 길이를 맞춰주는 과정이 필요한데, 사전에 지정한 max_length를 기준으로 입력 token의 길이가 길면 truncation, 짧으면 padding.

enc_ids) encoder의 입력으로 사용되는 feature.
enc_mask) encoder의 입력으로 사용되는 feature.
enc_ids에서 유효한 (padding이 아닌) 부분만큼 1의 값을 가지고 나머지는 0으로 채워짐.
dec_ids) decoder의 입력으로 사용되는 feature.
요약문을 생성할 수 있게 첫 값은 bos_token_id로 시작.
dec_mask) decoder의 입력으로 사용되는 feature. enc_mask와 동일한 역할.
label_ids) decoder의 label이 되는 feature.

2. Transformer (transformer.py)

- Transformer를 구성하는 요소 구현

  - MultiHeadAttention

    : Transformer의 multi-head attention 구현.

  - EncoderLayer

    : 구현한 multi-head attention을 바탕으로 Transformer Encoder layer 구현.

  - DecoderLayer

    : 구현한 multi-head attention을 바탕으로 Transformer Decoder layer 구현.

2. Transformer (transformer.py)

- Transformer를 이용해 모델 학습, 요약문 생성 코드 구현

  - forward

    : Transformer 학습에 사용되는 cross entropy loss 값을 계산하는 코드 구현.

  - generate

    : 학습된 모델을 이용해 enc_ids 와 enc_mask를 입력하는 해당 입력에 맞는 summary를 생성하는 코드 구현.

- Reference

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

3. Training, Evaluation (main.py)

- 학습 및 평가

  - configuration을 조정하며 모델 학습 및 평가

  - 평가 지표는 문서 요약 평가에 주로 사용되는 ROUGE-F1 score 사용

  - 학습과 평가를 끝낸 후 submit 데이터에서 만들어지는 요약문을 txt 파일로 제출

- Reference

- Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*(pp. 74-81).

채점 기준

- 채점 기준

  - 코드 실행 가능 여부

  - Transformer, 전처리 코드 구현

  - Submit 데이터에서 만들어진 요약문(hypothesis)과 실제 요약문(golden summary)을 이용해 ROUGE score를 측정했을 때,
    ROUGE-1 31.0 이상, ROUGE-2 20.0 이상, ROUGE-L 30.0 이상.

주의 및 안내사항

- 주의 및 안내사항
  - 제공하는 skeleton code 에서 Edit 부분 이외는 수정하지 말아주세요.
  - 제공하는 skeleton code 에서 import 한 library 이외의 library 는 불허합니다.
  - 채점은 Google Colab GPU 환경에서 진행할 예정입니다. Colab 환경에서 코드가 실행되지 않으면 0점 처리됩니다. 자세한 사용법은 아래 링크를 참고해주세요. (https://yjs-program.tistory.com/124)
  - Colab 환경에서 huggingface tokenizer와 rouge-score metric library를 사용하려면 아래와 같이 transformers, rouge-score 를 install 해주면 됩니다.

```
!pip install rouge-score transformers
```

주의 및 안내사항

- 주의 및 안내사항

  - 모든 과제는 시스템에 의해 copy 검사를 진행중입니다.

  - 본인이 수정한 3개의 .py 파일 (datamodule.py, transformer.py, main.py) 과 1개의 .txt 파일 (submit.txt) 을 압축한 후 '2022123456_홍길동.zip' 으로 파일명을 변경해서 제출 부탁드립니다.

  - 5월 20일 금요일 19:00 에 한시간 정도 과제 내용에 대한 QA를 진행할 예정입니다. 문의 내용은 해당 시간을 이용해 주시면 감사하겠습니다. 자세한 내용은 icampus 공지 확인 부탁드립니다.

  - 제출 기한 : 5월 29일 (일) 23:59