Week1-4 과제

▼ Week4 자료

NLU - STS(Semantic Textual Similarity) task

- 참고 논문: KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding
- 문제 정의
  - STS는 두 문장간의 의미론적 유사성을 평가하는 테스크이다. 두 문장의 유사성을 0과 5 사이의 값으로 표현하거나 0과 1 바이너리 값(STS-B)으로 표현한다.
- 데이터 셋 소개
  - KorSTS (한국어)
    - 소개: GLUE의 영문 STS 데이터를 한국어로 번역한 데이터
    - 입력 데이터는 두 개의 문장이며 출력 값은 0과 5사이의 값이(regression)거나 0또는 1의 값(binary classification)이다.
    - 학습 데이터 : 5,749
    - 검증 데이터: 1,500
    - 테스트 데이터: 1,379
- 모델 소개
  - 학습 방법론
    - cross-encoding
      - 문장 2개를 pair input 형식으로 모델에 입력
      - 성능이 우수하지만 연산량이 많아 학습 시간과 자원이 많이 필요
    - bi-encoding
      - 문장 2개를 single input 형식으로 모델에 개별적으로 입력
    - 성능은 cross-encoding이 bi-encoding보다 높음
  - 모델
    - RoBERTa
      - 대규모 코퍼스에 pre-train 한 모델을 가져와 KorSTS 데이터로 fine-tuning.
        - pre-training 데이터 셋
          - 한국어 corpora (65GB)
        - tokenizer
          - SentencePiece (vocab size: 32,000)
    - XLM-R
      - RoBERTa와 동일한 아키텍처
      - pre-training 데이터 셋
        - 다국어 코퍼스 (2.5TB)
      - vocab
        - vocab size: 250,000
- 평가 지표

    - STS linear regression metrics
        - Pearson Correlation Coefficient
        - 모델의 예측 값(y)과 실제 값(y)의 pearson correlation coefficient (r)는 -1과 1사이의 값으로, 평가 지표는 pearson correlation coefficient에 100을 곱한 백분율로 표현
        - Pearson correlation coefficient 설명 영상
    - STS binary classification
        - Accuracy (%)
            - TP + TNP + N100
            - 모든 테스트 데이터 중 실제로 긍정 또는 부정 레이블을 맞춘 경우를 백분율로 표현
- 성능 (내림차순)
    - RoBERTa (large) : 85.27%
    - XLM-R (large) : 84.68 %
    - RoBERTa (base) : 83 %
    - XLM-R (base) : 77.78 %

NLG - Extractive Summarization task

- 참고 논문
    - MATCHSUM
- 문제 정의
    - 문서 내에서 문서를 가장 잘 요약하는 단어나 문장을 찾는 테스크이다.
- 데이터 셋 소개
    - CNN/Daily Mail (영어)
        - 소개: CNN과 Daily Mail 온라인 뉴스 기사 코퍼스로 사람이 뽑은 요약 문장이 답으로 제시되어 있다.
        - 학습 데이터: 287,226
        - 검증 데이터: 13,368
        - 테스트 데이터: 11,490
- 모델 소개
    - 학습 방법론
        - summary-level framework
            - 기존의 sentence-level framework는 문장끼리 비교해 요약 문장을 찾는 방식이었지만 해당 framework는 문장 벡터와 문서 벡터간의 거리를 계산함
        - contrastive learning
            - 문서와 요약 문장간의 거리는 가까워지고 요약 문장이 아닌 문장과의 거리는 멀어지도록 학습
    - 모델
        - Siamese-BERT
            - 샴 쌍둥이처럼 2 개의 BERT 모델이 동일한 weight를 공유하고 있는 아키텍처를 의미함. 하나의 BERT 모델은 문서를 입력받고 다른 BERT 모델은 요약 후보군을 입력 받음. BERT 모델의 마지막 layer의 “[CLS]” 토큰을 추출해 입력 문서의 embedding으로 사용함.
        - loss function: margin-based triplet loss
            - Triplet loss는 문서(Document), 실제 요약 문서(Gold Summary), 그리고 요약이 아닌 문서(Candidate Summary)간의 상대적인 거리를 학습함. 문서와 실제 요약 문서의 거리는 가까워지고 문서와 요약이 아닌 문

서 사이의 거리는 멀어지도록 학습.

        - 추론 과정
            - 문서 embedding과 요약 문서 후보군들의 embedding간 cosine similarity를 계산 해 가장 높은 유사도를 갖는 요약 문서를 선택
- 평가 지표
    - ROUGE-1
        - 실제 요약 문장(정답) 중에서 모델이 요약한 문장(예측값)과 겹치는 단어의 비율을 측정

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

        - What is ROUGE and how it works for evaluation of summarization tasks? | RxNLP
            - 한국어 번역본
- 성능 (ROUGE-1)
    - BERT-base: 44.22
    - RoBERTa-base: 44.41

Week1-4 과제 리뷰

KorNLI and KorSTS 한국어 논문을 읽은 적이 있는데 cross-encoding과 bi-encoding 차이를 이해하는데 어려움이 있었는데 Week1-4 과제를 통해 그 차이를 알 수 있게 되었습니다. cross-encoding은 문장 2개 쌍을 입력하기에 연산량이 늘어나지만 우수한 성능을 보이고 bi-encoding은 문장 2개를 단일 입력하기에 연산량이 적을 것으로 예상됩니다. 또한 KLUE의 한국어 데이터도 그렇고 카카오브레인의 KorNLI, KorSTS 데이터도 그렇고 전반적으로 RoBERTa가 우수한 성능을 보이는 것으로 알고 있습니다.

추출 요약의 학습 방법론은 summary-level framework와 constrative learning이 있는데 전자는 문장 벡터와 문서 벡터 간의 거리를 계산하고 후자는 문서와 요약 문장 간의 거리는 가깝게 아닌 문장과의 거리는 멀어지도록 학습한다고 합니다. 샴 버트는 2 스테이지 모델이라고 볼 수도 있을 것 같습니다. triplet loss는 얼굴 인식하는 모델의 loss로도 자주 사용되는 것으로 알고 있습니다. ROUGE-1은 unigram recall이고 예측한 문장 중에 정답 문장에 속하는 개수의 비율로 스코어를 매깁니다.