기업과제3_10팀_권예환_개인보고서

선택한 모델

- Sentence-BERT 사용
  - 두 문장을 base 모델을 활용해 문장 임베딩 벡터을 추출하고, 두 벡터의 Cosine similarity를 loss로 사용하여 학습을 진행하였다.
  - 다른 시도했던 모델(koBERT, BERT) 보다 초기 성능이 높았고, 모델 구조상 real-label을 학습하기 좋아 STS 문제를 해결하기 좋다고 판단하여 이 모델을 사용하였다.
  - RoBerta 모델을 base로 사용하였고, base 모델의 output을 mean pooling 하여 문장 임베딩을 추출하였다.

추가로 시도한 방법

- 학습 데이터셋인 KLUE-STS가 이미 전처리가 잘 되어있다 판단하여 전처리 작업은 진행하지 않았다.
- Random swapping을 통한 데이터 증강을 시도했으나 쓰기 전과 비교해 성능이 높아지지 않아 해당 방법은 적용하지 않았다.

학습 파라미터

- Train dataset과 valid dataset을 9:1로 random split 하였다.
- epoch은 4로 하였다. epoch을 4로 하여도 성능이 높아 그대로 사용하였다.
- 훈련 배치 수의 10%를 warmup_step으로 하였다.

```
model.fit(
    train_objectives = [(train_dataloader, train_loss)],
    evaluator = valid_evaluator,
    epochs = epochs,
    evaluation_steps = int(len(train_dataloader)*0.1),
    warmup_steps = warmup_steps,
    output_path = model_path
)
```

학습 결과

- valid score : Pearson 0.9614, Spearman 0.9284
- dev score : Pearson 0.8887, Spearman 0.8889

담당한 역할

- Sentence-BERT 논문 리서치 및 분석하여 팀원에게 모델 사용 권유