기업과제 4. 학습 보고서

1. 자신이 담당한 역할 (ex. 모델링, 논문 리서치, 하이퍼파라미터 튜닝 등) 설명

- 논문 서치
- 추가데이터 크롤링
- 데이터 전처리 및 EDA
- 학습 및 평가
- wandb를 이용한 하이퍼파라미터 튜닝
- 평가지표 연구 및 설정
- Metric 모듈 구현

모든 과정을 다른 팀 멤버들의 피드백을 받으며 주도적으로 완성하였다.(팀 전체인원이 3명이었기에, 본인이 과제4를 담당함)

2. "TRUE SUMMARY" 생성한 방식 및 근거

- Extractive Summary가 아닌 abstractive summarization 모델을 선정하였고, pre-trained model을 불러와서 사용했기 때문에, TITLE을 true summary로 학습하여 기존의 성능 대비 **'기사 제목스러운'** 요약문을 생성할 수 있도록 목표를 설정하였다. 이에 따라 TRUE SUMMARY로 주어진 Dataset의 TITLE을 그대로 이용하였다.

3. 평가 지표 함수를 선택 또는 정의한 근거

A. Ideation

i. 요약 모델의 평가지표는 형태적 유사성(Rouge)과 의미적 유사성(Embedding similarity) 이 모두 고려되어야 한다고 생각하였다.

ii. 특히 한국어는 어순보다 조사가 우선되는 특징, 그리고 단어의 변형이 심하다는 특징 등에 의해 Rouge score로 평가하기에는 분명한 한계점이 있다고 판단하였다.

> **ex) 한국어의 차별화된 특징**
>
> 1. 어순보다 접사가 중요하다.
> 아빠가 나에게 선물을 주셨다. == 나에게 아빠가 선물을 주셨다.
> → Daddy bought me the gift. != I bought Daddy the gift.
>
> 2. 단어의 변형이 많다.
> (보내, 보낼, 보냈...) 모두 같은 의미이지만,

ROUGE score 집계 시 모두 다른 단어로 인식

B. 평가 지표 함수의 구성(metrics.py 함수 참조)

i. **metric_embed** : SentenceTransformer pretrained model을 활용한 유사도

- 한국어 dataset으로 pretrained 된 문장 유사도 측정 모델을 활용하여, **제목과의 유사도 뿐 아니라 본문과의 코사인 유사도를 측정**하여 평균값으로 취함.

ii. **metric_rouge** : 불용어 및 특수문자 제거 후 형태소 기준 rouge-1 F1 score

- clean_text : 특수문자에 대한 전처리 진행
- crawl_stopwords : 한국어 불용어 리스트 크롤링
- Rouge : Unigram, bigram 및 skip-bigram의 attribute를 생성하는 Rouge class를 구현함. 이 때, 불용어 리스트를 활용하여 불용어 제거.
- Text_to_gram_instance : konlpy 를 활용하여 형태소로 분리 후 Rouge 이용하여 인스턴스 생성
- Scores : output과 title을 비교하여 precision, recall 그리고 f1 score를 계산

iii. **metric** : metric_embed 와 metric_rouge의 평균값이자, 성능 최종 평가 지표

C. metric 정의에 대한 근거

i. 형태적 유사성 : 왜 Rouge-1 F1 score인가?

- Rouge score에는 대표적으로 Rouge-1(Unigram), Rouge-2(bigram) 그리고 Rouge-L(LCS) score로 측정된다. 단어의 순서가 특히 중요한 Rouge-2나 Rouge-L같은 경우 한국어의 특성상 순서 기준으로 지표를 설정하는 것이 적절치 않다고 판단하였다. 어순보다 조사의 적절성이 더 중요하기 때문이다.
- 더불어 Rouge-1 score는 평균적으로 rouge-2 score보다 두 배 이상 높게 측정된다. Rouge-2 score로 의미적 유사성도 함께 고려하는(평균을 내는) 작업에 있어 embedding score와의 편차를 더 크게 만들 것이라고 판단하였고, 이는 score의 평균치를 낮추어 metric의 분포의 최대, 최소값을 작게 만들고, score 분포를 더 유의미하게 확인하지 못할 것이라고 판단하였다.
- Output을 기준으로 recall만 확인하기 보다는, title의 중요 형태소(키워드)를 기준으로 판단하는 precision도 함께 고려하여 F1-score를 선정하였다.

ii. 의미적 유사성 : 왜 SentenceTransformer인가?

- Embedding vector를 본문, 제목, output으로부터 추출하여 의미적 유사도를 비교할 수 있으면 좋겠다고 생각하였고, 한글 pretrained model을 조사하였다.
- 그 중 단어(Word2Vec, Fasttext)가 아닌 문장을 기준으로 비교하여 cosine similarity를 구하는 모델을 선정하게 되었고, 이는 vector 간 각도를 기준으로 유사도를 측정하기 때문에 content와 output간의 길이 차이가 많이 나도 같은 차원의 embedding vector로 매핑하여 비교한다. 이러한 점을 주목하여 metric을 고안하였다.

4. 모델 선정 배경 및 이유

A. 선정 기준

i. 한국어 요약본 dataset으로 Pretrained 된 모델인가?

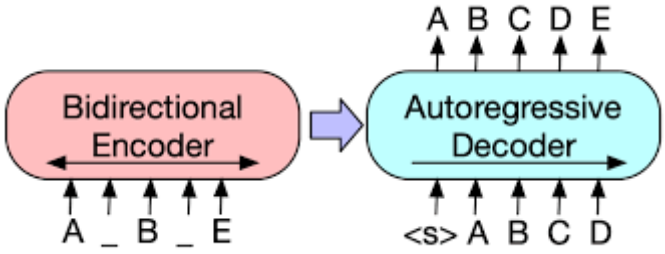
ii. 영어 NLP Abstractive Summarization 중 상위에 랭크된 SOTA 모델인가?

iii. 모델을 불러오고 이용하기 편리한가?

B. 모델 선정 과정 및 이유

i. Abstractive summarization task로 한국어 요약 dataset이 pretrained 된 모델을 조사하였고, 특히 기사 형태의 데이터셋을 학습한 모델을 목표로 검색한 결과 kobart-news 모델로 최종 선정하였다.

ii. BART는 MLM을 활용한 bi-directional BERT model과, 단방향 autoregressive model인 GPT를 결합한 모델이기에 요약 task에 적용되기에 적절한 모델이다.



iii. BART는 abstractive text summarization에서 SOTA 성능을 보이며, KOBART 는 한국어 데이터로 학습된 pre-trained model이다. 그 중 본 모델은 뉴스기사 요약 데이터셋을 학습시켰으므로 본 과제에서 더 성공적일 것이라 판단하였다.

iv. 해당 모델은 huggingface library를 통해 이용할 수 있었기 때문에 짧은 시간 안에 모델의 성능을 확인하고 검증할 수 있었고, 학습 전 요약 성능을 정성 분석한 결과 보기 좋게 요약되는 것을 확인할 수 있었다.

5. 모델 훈련 과정

A. 추가데이터 크롤링

i. Validation을 위한 1000개의 dataset을 추가 크롤링하였다. 기존 데이터와의 일관성을 위하여 같은 매거진(에서 추출되었다고 판단되는)의 데이터를 3월 20일자 기준 최근 data 1000개(본문, 제목 set) 크롤링하였다. (Beautifulsoup 이용)

B. 데이터 전처리 및 EDA

i. 본문 및 제목 길이, 요약에 있어 필요 없다고 판단되는 부분(기자 이름, 매거진 이름, html 태그 등)을 제거하였다.

ii. 분포를 확인하고 GPU error 방지를 위해 너무 긴 본문 data와 같은 경우 truncation을 거쳤다.

iii. Dataset을 구성하고, train, validation, test를 위한 Dataloader를 구성하였다. Dataloader는 collate function으로 tokenizing을 및 batch화, sampling을거쳤다.

C. Training & Validation

1. Loss : CrossentropyLoss
2. Optimizer : AdamW
3. Scheduler : get_linear_schedule_with_warmup

ii. Train함수(이후 sweep함수)와 validate 함수를 정의하고, epoch 별로 번갈아 가며 검증할 수 있도록 학습을 진행하였다. 모델의 Cross entropy loss를 최적화하는 방향으로 학습을 진행하였다. 다시 말해, training loss를 batch 별로 측정하고 validation set에서의 평균 loss를 통해 학습 추이를 판단하였다.

iii. Loss는 model이 label 입력 시 output으로 자동으로 반환해 주기 때문에 별도로 정의하지 않았다.

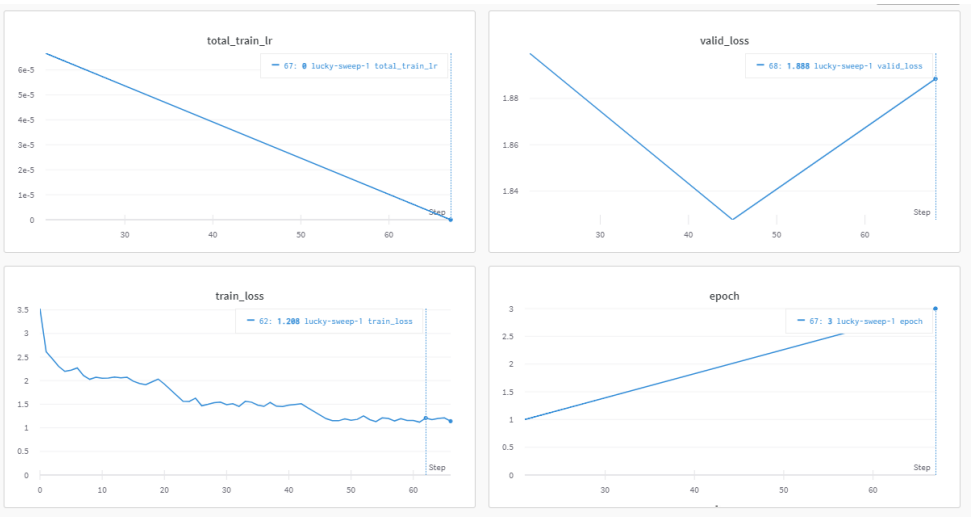
6. 모델 튜닝 과정

A. Training : wandb를 이용한 hyperparameter tuning 및 optimization

i. 최종 선정된 hyperparameter(validation loss : 1.8276, Bayesian optimization)

batch size	learning rate	weight_decay	epoch	val_loss
16	1.00E-05	0	1	2.21
			2	2.11
			3	2.13
32	1.00E-04	0.1	1	1.9
			2	**1.83**
			3	1.89
			1	1.89
			2	1.83
			3	1.86
	1.00E-05	0.1	1	2.34
			2	2.17
			3	2.17
		0	1	2.29
			2	2.15
			3	2.18
final val_loss				**1.8276**

1. **Epoch : 2 (대체로 3이상일 시 val_loss가 하락하는 모습을 보임)**
2. **Learning rate : 1e-4**
3. **Train_batch_size : 32**
4. **Weight_decay : 0.1**
5. **Warm_up_ratio : 0**



[Epoch 3번 간 training loss, learning rate, validation loss 변화]

ii. Epoch, learning rate, batch size, weight decay(Regularization), warm_up ratio 등을 고려하여 validation set의 loss를 최소화하는 과정을 진행하였다.

iii. Bayesian optimization을 활용하였기 때문에, prior loss 값을 고려하여 parameter를 선

택하였다. 때문에 모든 config 값을 훑지 않았다는 것이 grid search와 중요한 차이점이다.

iv. Validation loss는 여러 번의 epoch와 값들을 실험해본 결과 1.9~2.1 의 validation loss를 웃도는 것을 확인할 수 있었고, 최종적으로 1.8276의 validation loss를 보이는 모델로 선택하였다.

v. Weight decay는 bias 및 layer normalization에는 적용하면 안 되므로, 따로 적용되지 않도록 설정해주었다.

B. Hyperparameter tuning with Generation

i. 최종 선정된 hyperparameter

metric score		num_beams		
		4	5	6
length penalty	0.5	0.481	0.4804	0.4829
	1	-	0.4848	0.4887
	2	0.4808	0.4884	**0.4939**
	3	0.4773	0.484	0.4883

1. **Num_beams : 6**
2. **Length_penalty : 2**

ii. 별도로 구성한 metric으로 score를 측정하며 test set에서 가장 높은 score를 보이는 hyperparameter로 채택하였다.

iii. Wandb를 이용하지 않고 단순한 코딩을 통한 grid search로 선정하였다.

iv. Beam 개수가 많아지면 모델일 선택해야 할 선택지가 늘어나므로 시간이 오래 걸린다. 값이 클수록 6이 가장 적절하다고 판단되었다.

v. Length penalty 값이 너무 높거나 낮으면 모델이 할 말을 찾지 못해 공백을 반복하거나 문장이 마무리되지 않는 현상이 나타났고, 2일 때 가장 적절하였다.

vi. 생성 요약문의 max length는 title과의 비교를 위하여 적절한 값을 별도로 설정하였다.

7. 최종 결과 분석

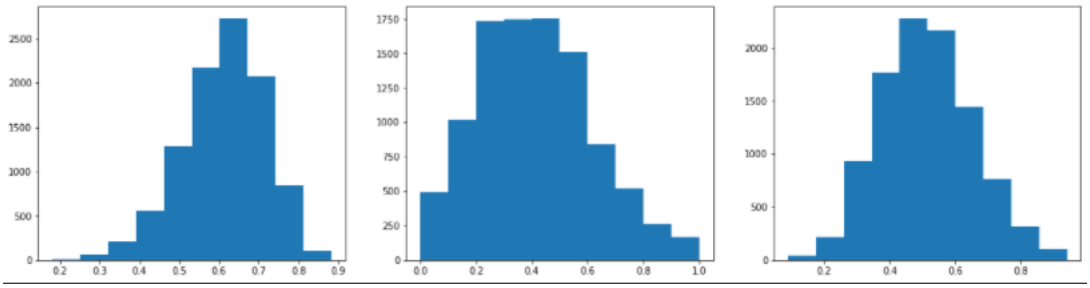
A. Test loss : 1.8573

Test set loss는 validation set loss와 큰 차이를 보이지 않았다. Training set과 Testset의 데이터 일관성 및 정규화(weight decay)가 그 이유인 것으로 판단된다.

B. TRUE SUMMARY, PREDICT SUMMARY, SCORE

	TRUE SUMMARY	PREDICT SUMMARY	SCORE
409	정신 차려 맨유, 호날두 대안이 '프랑스 9골 공격수'	'호날두 대제차' 맨유, '프랑스 9골 공격수' 호날두 대제차로 찜	0.594047
6420	'90분으로는 부족해' 리버풀, 클롭 체제 추가 시간 13골 가동중	'살라도 못 해!' 리버풀, 울버햄튼 원정서 오리기 결승포... 리버풀 선두 등극	0.340392
3424	박동혁 감독, "초반 실수로 흐름 뺏겼다...선수들 급해졌어"	박동혁 감독, "경기 초반 실수로 흐름 뺏겼다"...전남과 격차 좁히기 실패	0.630406
1984	'윤일록 데뷔골' 울산, K3 양주 돌풍 잠재웠다... 2-0 누르고 4강 진출	' 윤일록+김지현 멀티골' 울산, 양주 2-0 격파... 4년 만에 FA컵 우승 도전	0.573605
6604	탈락 위기인데 부상자 속출 밀란, 리버풀전 어쩌나	'더 최악' 밀란, 리버풀전 38회 결장...아틀레티코전도 결장	0.367062
556	'못 뛰어도 경기장에 있어주지...' 포르투갈로 떠난 호날두 향한 아쉬움	"날두 경기에 뛰지 못해도 경기장에 있어줘" 맨유 동료들의 아쉬움	0.479707
9391	'사실상 방출 명단' 입지 불안한 토트넘의 11명 많기도 하네	토트넘의 11명 중 가장 미래가 불확실한 라인업, 방출보다 임대가 더 중요	0.473941
480	컵스 마무리 후보, 직장폐쇄 최대 피해자 될까...토미 존 수술로 시즌 아웃	'장장폐쇄' 호이어, 토미 존 수술로 시즌 조기 종료... 선수와 컵스 모두에게 큰 악재	0.548579
8343	바르사 리턴 무산되자 프리미어리그 복귀설...에버튼 '역제의' 받아	'르르' 산체스, 이번엔 프리미어리그 복귀설...에버튼, 인터밀란에 역제안	0.589310
6372	'홀란드 데려오라'...에이전트 랑닉에게 거액 보너스?	랑닉, 맨유에서 연봉 800만 파운드 125억 받겠네...홀란드 영입하면 157억 ...	0.338429

C. Score(왼쪽부터 sim_score, mor_score, final_score)



i. 임베딩 기반 의미적 유사도는 평균 0.61, Rouge 기반 형태적 유사도는 평균 0.41의 score를 보였다. 이를 평균내니 t분포와 같은 비교적 대칭적인 분포를 보였다.(평균 0.51)

ii. Similarity score와 morpheme score 간에는 0.666 의 correlation coefficient 를 보인다. 이는 의미적으로도 유사하다면 형태적으로도 어느 정도 유사하다는 것을 의미하지만, colinearity를 고려할 만큼 강한 상관관계(1에 가까운)가 아니기 때문에 별도로 분리해서 측정할 만한 가치가 있다고 판단된다.

D. Training set/ test set과의 Two-sample T-test

i. 모델의 정규화 성능(test set과 training set 간의 score 차이)를 보기 위하여 two-sample t-test를 진행하였다.

ii. 귀무가설을 '두 샘플간 score 평균값에 차이가 있다' 로 놓고 각 통계량을 측정하고, t value 값을 도출한 후 z-score table에 의해 귀무가설을 부정하여 유의성을 확보하

였다.(95%의 Confidence)

iii. 이는 training set에 과적합 되지 않고 testset에도 잘 적용될 수 있음을 의미한다.

E. 평가지표에 대한 평가

i. 데이터를 추가로 수집하고, hyperparameter 를 tuning해감에 따라, 고안한 metric의 두 score(`sim_score` 와 `mor_score`)사이의 correlation이 점점 올라가는 것을 확인할 수 있었다. 또한, 의미적 유사도 점수에는 큰 변화가 없었지만 형태적 유사도 점수에는 상승을 확인하면서 Training loss를 최적화하는 과정이 결국 label과의 token 유사도와 관련이 있다는 것을 이해할 수 있었다.

ii. 평가지표 중 가장 높은 점수를 기록한 data 및 결과를 확인해 본 결과, 거의 똑같다고 볼 수 있는 요약본을 생성한 것을 확인하였다.

```
'[오피셜] 램파드, 에버턴 감독으로 EPL 복귀... 2024년까지 계약',
'램파드, 에버턴 감독으로 EPL 복귀...2024년까지 계약')
```

iii. 유사도 점수가 떨어지는 output으로 비교한 결과, 요약본이 본문의 의미와 전혀 다른 내용임을 확인하였다. 이를 통해 합리적인 평가지표를 고안했다는 것을 확인할 수 있었다.

```
' '부상 우려' 를 활발함과 예리함으로 씻어낸 황의조',
' "시차와 비행, 잠을 못 잔 게 원인이었어" 황의조, 보르도 최전방에 섰다')
```

F. 개선을 위해서 시도해볼 사항

i. 일부 summary에서 첫 단어를 여러 번 반복하거나 단어 표현에 어긋나는 다른 글자를 고르는 것을 확인 -> 이는 abstractive summary이자 autoregressive한 decoder에서 첫 단어를 고르는 데에 어려움이 있기 때문으로 판단됨. 추가 학습이나 모델 교체로 개선이 가능할 것으로 판단된다.

ii. 본질적으로 BART model은 decoder output과 label 간의 Cross entropy loss를 최적화하는 모델이기에, Training 시에는 validation loss를 최소화하는 metric, Generation에서는 rouge score를 최대화(loss값은 생성 시 hyperparameter값에 영향을 받지 않음)하는 방향으로 hyperparameter를 조정하였다. 이후 설정한 평가지표의 점수 추이도 함께 관찰하며 학습시키는 과정도 관찰하면 또 다른 흥미로운 결과를 볼 수 있을 것으로 예상된다.