Goorm Project#2

Machine Reading Comprehension

< 2조 (2조) >

조현수, 이준엽, 김남준, 김영수, 손동협, 조승연, 정명관

01 PPT PRESENTATION

01 프로젝트 개요

- 프로젝트 개요 및 개발환경
- 프로젝트 진행과정
- 기대효과

02 프로젝트 팀 구성 및 역할

- 프로젝트 팀 구성 및 담당 part

03 프로젝트 진행 프로세스

- 전체일정

04 프로젝트 결과

- 결과 및 도출 과정

05 자체 평가 및 피드백

- 평가 및 보완점
- 조원 간의 피드백

06 Paper Reference

Q&A

01 프로젝트 개요

01 프로젝트 개요 및 개발환경

프로젝트 구현 내용

한국어 지문을 보고 질문에 맞는 답을 추출하는 모델을 만들고 성능을 개선

활용 라이브러리, 프레임워크, 모델

Huggingface Transformers
- KoELECTRA
- KoBigBird

WandB

Sweep

01 프로젝트 개요

02 프로젝트 진행 과정

베이스라인 코드 정리

데이터 전처리 및 분석

모델 선정 및 다양한 기법 적용

모델 학습

WandB분석

Hyper-parameter tuning

Submission 제출

01 프로젝트 개요

03 프로젝트 기대효과

1. 너무 긴 문서에 대해 필요한 정보를 효율적으로 추출

2. 사용 도메인에 맞는 추가적인 데이터를 이용해 모델을 해당 도메인에 적용

3. 스마트 스피커나 챗봇 등으로 확장

02 프로젝트 팀 구성 및 역할

01 프로젝트 팀 구성 및 담당 part

훈련생	역활	담당업무
조현수	팀장	- 베이스라인 코드 정리 및 코드 수정, 전체part 관리 - Model 탐색 및 선정
김남준	팀원	- 여러 pretrained 모델 탐색 및 적용 - 데이터 분석 및 전처리
김영수	팀원	- learning rate scheduler, gradient clipping 적용 및 분석 - 모델 파인튜닝 및 성능 개선
손동협	팀원	- Input Sequence 시각화, 입력 데이터 분석 및 전처리 - 모델 파인튜닝 및 성능 개선
이준엽	팀원	- 추가 학습 데이터 수집 및 분석, - 레벤슈타인 거리 코드 작성 및 평가지표로 사용
정명관	팀원	- huggingface의 fine-tuned 모델 탐색 및 실험 - WandB를 이용한 learning rate별 성능 분석
조승연	팀원	- fine-tuned 모델 탐색 및 실험 - 데이터 분석 및 전처리

02 프로젝트 팀 구성 및 역할

01 프로젝트 팀 구성 및 담당 part

공통part
1. WandB를 활용한 데이터 분석 및 실험결과 정리 2. 코드 분석 및 정리 3. 모델 fine-tuning 및 성능 개선 4. Learning rate 선정 5. 발표자료 정리

Fine-tuning 및 데이터 정리, 분석
1. Learning rate scheduler 분석 및 선정 2. WandB Sweep을 통한 최적의 hyper-parameter 탐색

모델 성능 개선 part
1. Ai 허브에서 데이터 추가 및 데이터 분석 2. train, valid dataset 비율 조정 및 병합 3. Optimizer들 분석 및 선정 4. 레벤슈타인 거리 코드 작성 및 평가 지표로 사용 5. 데이터 preprocessing 6. 앙상블 기법 적용 (성능이 잘 나온 모델 위주로 적용) 7. 결과데이터 시각화 8. 회의 내용 정리

03 프로젝트 진행 프로세스

01 전체일정

두 번째 프로젝트 : Goorm NLP project#2 - Machine Reading Comprehension

- 4/04 프로젝트 공개
- 4/04 ~ 4/18 총 15일 간 진행
- 주제 선정부터 보고서 작성까지 총 5단계에 걸쳐 프로젝트 진행

구분	기간	활동	비고
주제 선정 및 기획	4/04 (월)	- 프로젝트 고지 - Kaggle Competition 공개	-
모델 선정	4/05(화) ~ 4/09(토)	- BERT, KoELECTRA, KoBigBird 등 여러가지 모델 실험 및 선정	-
모델 튜닝	4/05(화) ~ 4/16(토)	- Learning rate, random seed, input sequence length, epoch 등의 hyper-parameter tuning	-
모델 성능 개선	4/07(목) ~ 4/16(토)	- Ensemble method 적용 - Probability distribution을 이용한 token 위치 및 개수 결정 - Input sequence length 조정 - 추가 학습 데이터 수집 - 데이터 전처리 등	-
자료 수집 및 보고서 작성	4/16(토) ~ 4/18(월)	- WandB의 그래프 및 모델 성능 비교 표 작성	-
총 소요기간	15일	-	-

04 프로젝트 결과

01 결과 및 도출 과정

Goal: Minimize Levenshtein Distance

Result

Model	Dataset	Dataset Size	Epoch	Learning Rate	Batch Size	Max Input Sequence	Max Output Tokens	Levenshtein Distance
KoBigBird	SQuAD1.0 + AIHub	355,730	2	6e-5	256 accumulate 64	1024	8	1.994(public) 2.011(valid)

04 프로젝트 결과

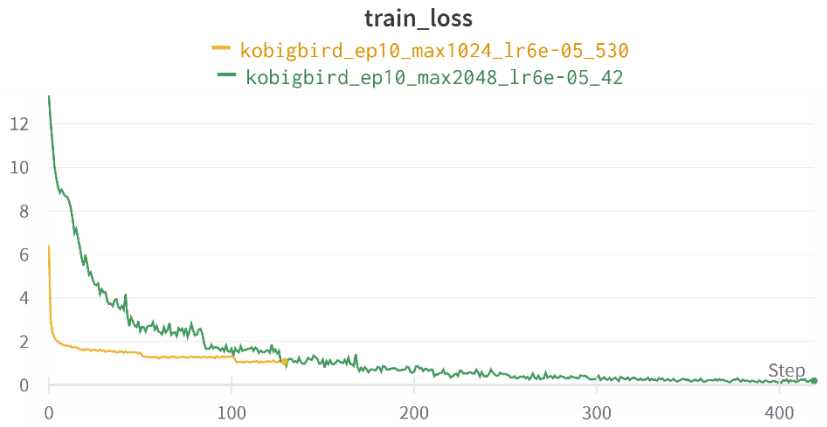
01 결과 및 도출 과정

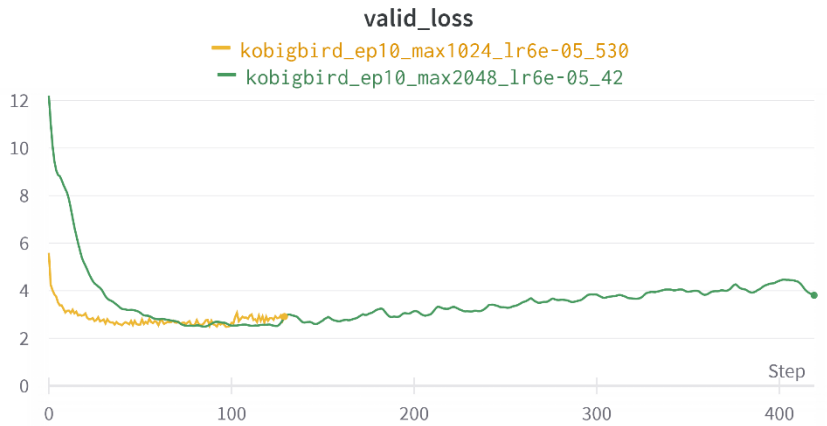
Input Sequence Information

	Max Length	Min Length	Mean Length
Train	1,187	248	552
Valid	1,165	256	556

Model Information

	Max Token Length
BERT	512
ELECTRA	512
Longformer	4,096
BigBird	4,096





best performance when 1024 input tokens

04 프로젝트 결과

01 결과 및 도출 과정

Datasets

Total	Train	Valid
355,730	350,000	5,730

{'answers': [{'answer_start': 0, 'text': '한국청소년단체협의회와 여성가족부'}],
'context': "한국청소년단체협의회와 여성가족부는 22일부터 28일까지 ...
'guid': '0e8d5fdb267d4130a5e85ef2cea9895e',
'question': "서울과 충북 괴산에서 '국제청소년포럼'을 여는 곳은?"}

{'answers': [{'answer_start': 19, 'text': '22일부터 28일'}],
'context': "한국청소년단체협의회와 여성가족부는 22일부터 28일까지
'guid': '5d93b1b2f3c54b0ebdae07a60e68101d',
'question': "'국제 청소년포럼'이 열리는 때는?"}

데이터셋명	기계독해			
데이터 분야	음성/자연어	데이터 유형		텍스트
구축기관	마인즈랩	데이터 관련 문의처	담당자명	안준환(마인즈랩)
가공기관			전화번호	031-625-4349
검수기관			이메일	pworks@mindslab.ai
구축 데이터량	45만	구축년도		2018년
버전	1.0	최종수정일자		2019.05.15
소개	기계독해 개발에 활용될 수 있는 뉴스 본문 기반 학습 데이터셋 45만 건을 구축한 지식베이스 제공			
주요 키워드	뉴스 본문 데이터셋, 정답 없는 데이터셋, 설명 가능 데이터셋, 표준 데이터셋, 질문, 단서, 답, 기계독해			
저작권 및 이용정책	본 데이터는 과학기술정보통신부가 주관하고 한국지능정보사회진흥원이 지원하는 '인공지능 학습용 데이터 구축사업'으로 구축된 데이터입니다. [데이터 이용정책 상세보기]			

데이터셋명	일반상식			
데이터 분야	음성/자연어	데이터 유형		텍스트
구축기관	마인즈랩	데이터 관련 문의처	담당자명	안준환(마인즈랩)
가공기관			전화번호	031-625-4349
검수기관			이메일	pworks@mindslab.ai
구축 데이터량	15만	구축년도		2017년
버전	1.0	최종수정일자		2018.01.02
소개	한국어 위키백과 내 주요 문서 15만 개에 포함된 지식을 추출하여 객체(entity), 속성(attribute), 값(value)을 갖는 트리플 형식의 데이터 75만 개를 구축한 지식베이스 제공.			
주요 키워드	한국어, 위키백과, 일반상식, 지식베이스, WIKI 본문, 질의응답, 챗봇, 지능형 QA 서비스, 위키백과 데이터			
저작권 및 이용정책	본 데이터는 과학기술정보통신부가 주관하고 한국지능정보사회진흥원이 지원하는 '인공지능 학습용 데이터 구축사업'으로 구축된 데이터입니다. [데이터 이용정책 상세보기]			

본 데이터는 과학기술정보통신부가 주관하고 한국지능정보사회진흥원이 지원하는 ' 인공지능 학습용 데이터 구축사업 ' 으로 구축된 데이터입니다.

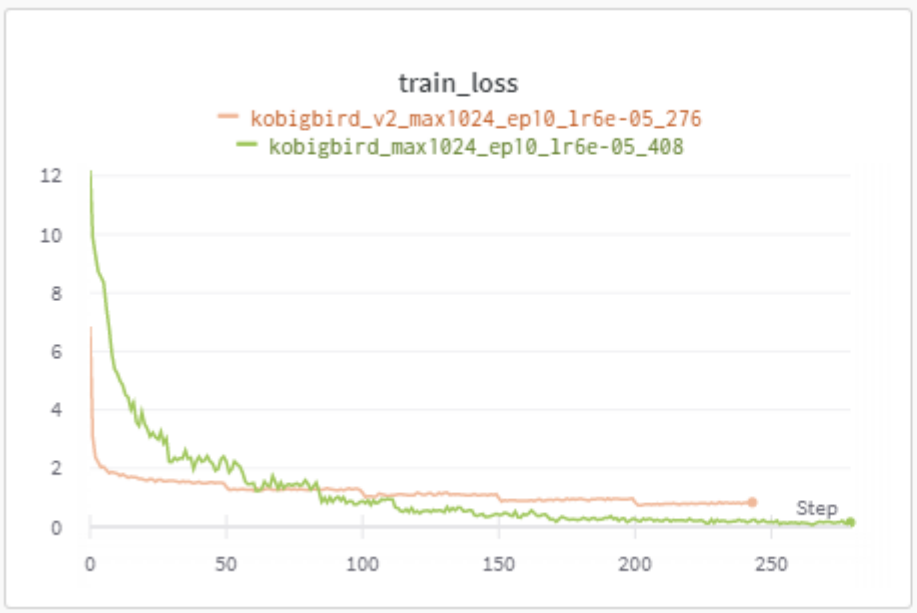
Ref: https://aihub.or.kr/aidata/86, https://aihub.or.kr/aidata/84

04 프로젝트 결과

01 결과 및 도출 과정

Datasets

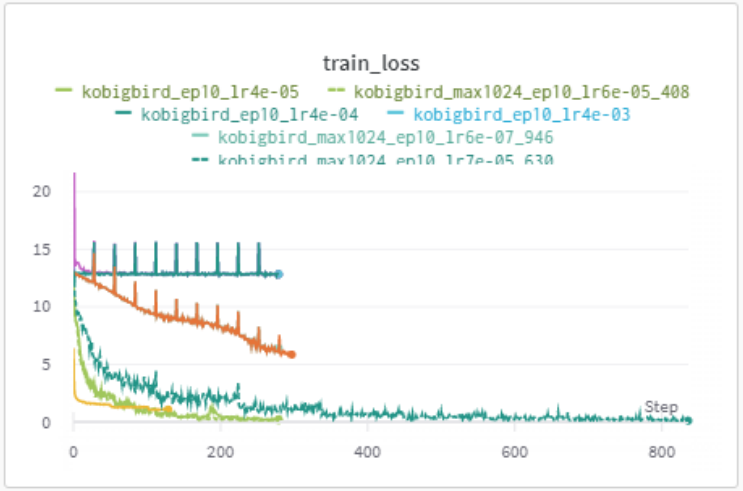
Total	Train	Valid
355,730	350,000	5,730

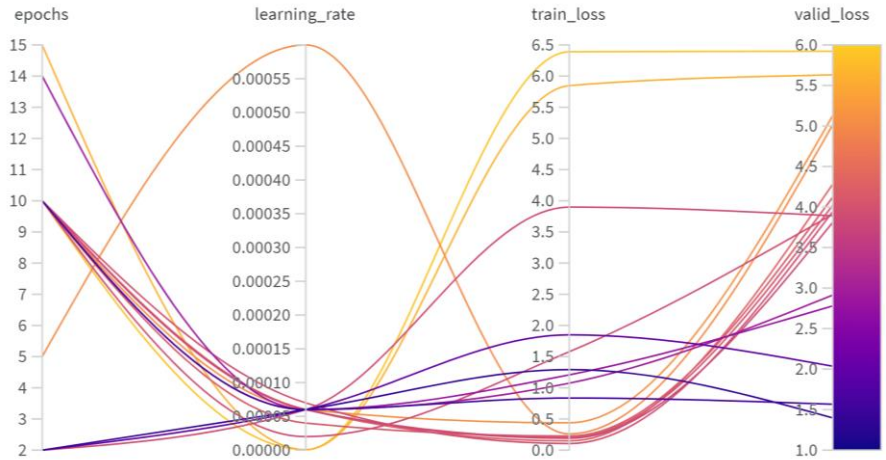


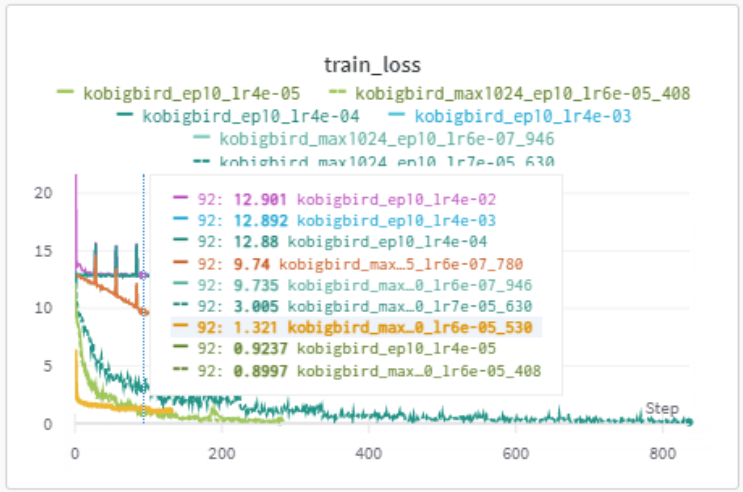
04 프로젝트 결과

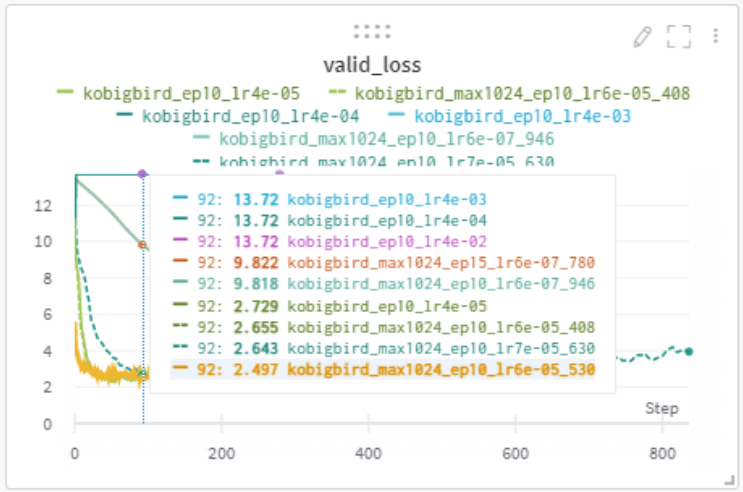
01 결과 및 도출 과정

Learning Rate Tuning







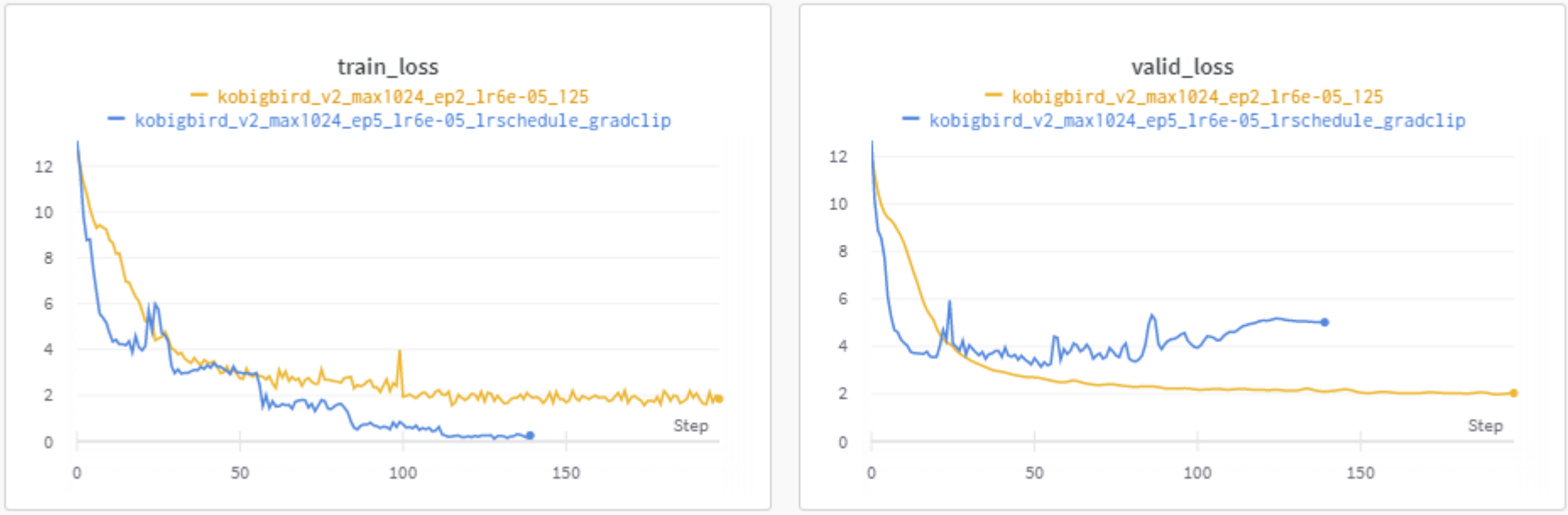


Learning Rate: 6e-5

04 프로젝트 결과

01 결과 및 도출 과정

Gradient Clipping + Learning Rate Scheduler

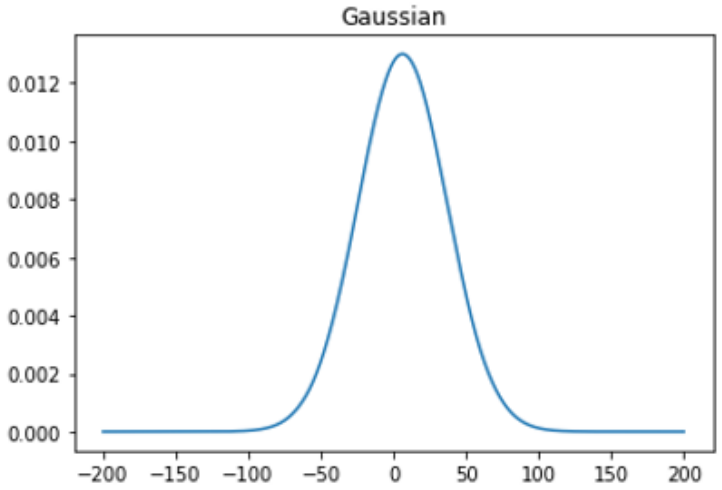


04 프로젝트 결과

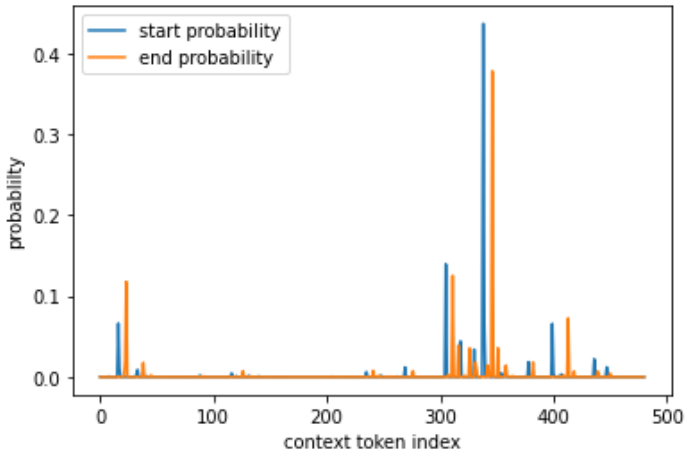
01 결과 및 도출 과정

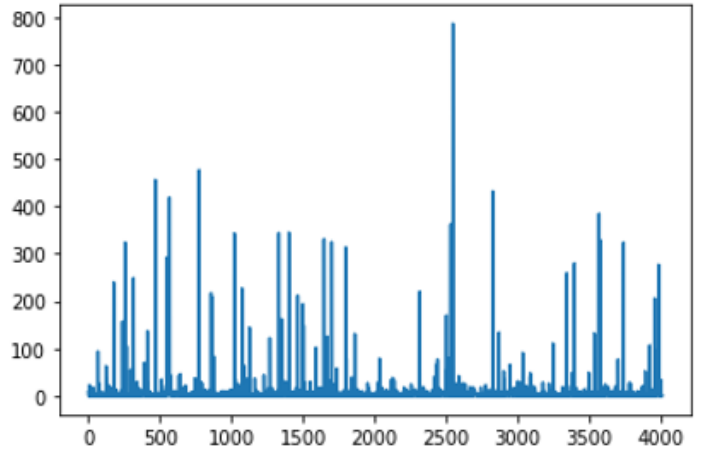
Output Token Length

Mean	Standard Deviation	Confidence Interval 90%	Confidence Interval 95%	Confidence Interval 99%
6.23	30.68	5.43 ~ 7.03	5.28 ~ 7.18	4.98 ~ 7.48



Token Length: 8





04 프로젝트 결과

01 결과 및 도출 과정

- Various random seed (42, 777, 1004)

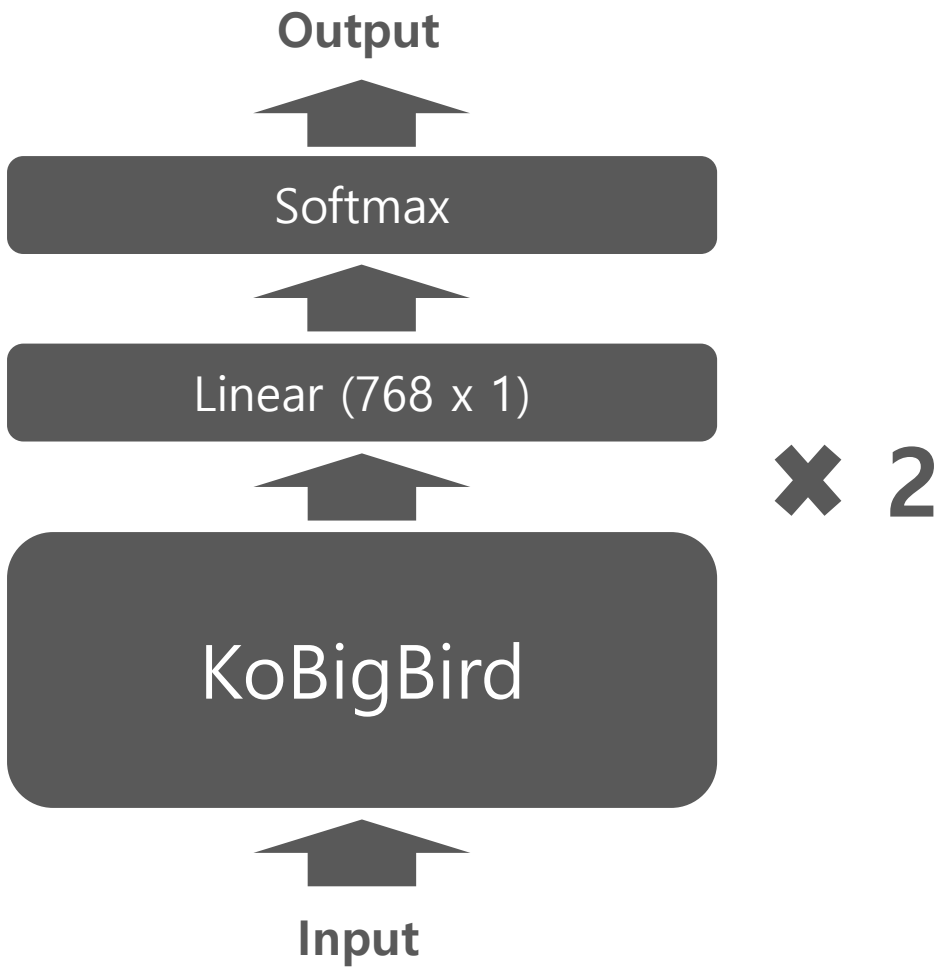
- Use Huggingface fine-tuned model

- Train with shortest answer

- Train model for start, end token for each

- Ensemble (mean token positions, mean string positions)

- Calculate Levenshtein Distance with valid dataset



05 자체 평가 및 피드백

01 평가 및 보완점

1. 데이터 및 결과를 여러 방법으로 분석하고 성능을 높일 수 있었다.
2. 베이스라인 코드 정리에 시간이 오래 걸렸다.
3. 한정된 GPU 자원으로 더욱 다양한 hyper parameter를 실험해보지 못했다.
4. 한정된 시간으로 다양한 Model을 사용해보지 못했다.
5. 성능 기록 및 자료 정리를 처음부터 잘 했으면 하는 아쉬움이 있다.
6. Model별 Tokenizer가 달라 전처리 하는데 어려움이 있었다.

05 자체 평가 및 피드백

02 조원 간의 피드백

조현수: 모델 성능 개선을 위해 데이터 전처리와 분석 부분이 중요하다는 것을 배울 수 있었고 모델을 직접 설계해보며 학습을 시켜보는 좋은 경험이었다. 기회가 된다면 huggingface 라이브러리를 사용하지 않고 새로운 모델을 구축하여 pretraining 부터 학습을 해보며 성능을 개선해보고 싶다.

손동협: 프로젝트기간이 길어서 단순히 모델과 하이퍼 파라미터만 바꾸는것만이 아닌 데이터 처리랑 내부구조를 이해하고 그에 따른 방법을 제시하는 등의 값진 시간이 되었습니다.
단순히 파라미터 조정하는것을 떠나 최선의 답을 내기 위한 생각을 하면서 팀원과 공유했던점이 매우 좋았습니다.

이준엽: 지난 프로젝트보다 시간적 여유가 조금은 생겨서 그래도 해볼 수 있는 것들이 많았다. 스스로 자원해 한 부분을 맡고 그것을 해결하는 것은 즐거웠다. 이게 팀프로젝트의 묘미인가. 하지만 코랩프로의 GPU Memory로 처리하기가 힘들 정도로 데이터를 쓰다보니 더 다양한 파라미터를 적용하지 못한 것은 아쉽다. 파라미터를 바꿔서 하더라도 큰 영향이 없어서 곤란했지만 후 처리로도 점수를 올릴 수 있다는 것을 알게됐다. 뭘 바꿔야 더 좋은 모델이 나올지는 예측이 가능했으나 그것을 뒷받침하는 근거를 찾는 것이 더 어려웠다.

김남준: 프로젝트를 진행하는 과정에서 다양한 모델을 찾아보고 이를 적용할 수 있을 지 검토하는 과정을 통해 더 넓은 시야를 가지고 판단할 수 있는 가능성을 엿볼 수 있었습니다. wandb에서 시각화시키고 추합하는 과정을 통해 시각화의 힘이 얼마나 대단한지 느낄 수 있었습니다

정명관: 전체적인 코드를 수행하는 시간이 오래걸려 다양한 모델 적용 및 실험을 못해봐서 아쉽습니다. 시간이 있다면 더 다양한 모델을 사용해보고 싶습니다. 그리고 전체적으로 적용된 코드들을 시각화자료로 보아 결과를 볼 수 있어서 이해하는데 도움이 되었습니다

김영수: 이번 프로젝트를 통해 프로젝트의 채점 방식, 결괏값 확인에 있어 다른 시야를 가지게 되었으며 제한을 걸어 진행하였기에 촉박한 시간과 모자란 gpu자원등을 효율적으로 사용하기위한 고민이 추가되어 발전에 큰 도움이 되었습니다

조승연: 단순히 파라미터 조정이 아니라 다양한 방법론과 모델을 사용하여 이번과제를 해결한 것이 좋은 경험이 되었습니다. 생각보다 다양한 방법들이 적용가능하고 좋은 결과가 나와 신기했다.

Paper Reference

BERT
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

ELECTRA
Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

Longformer
Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

BigBird
Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." Advances in Neural Information Processing Systems 33 (2020): 17283-17297.

Q & A

Thank You