

Upstage AI Lab

Dialogue Summarization 경진대회
NLP 1조

24.11.28

www.fastcampus.co.kr

Copyright © FAST CAMPUS Corp. All Rights Reserved. 무단전재 및 재배포 금지

목차

- 01. 팀원 소개
- 02. 대회 개요
- 03. 모델 선정
- 04. 성능 최적화
- 05. 결과 분석
- 06. Troubleshooting
- 07. 진행 소감

01

팀원 소개

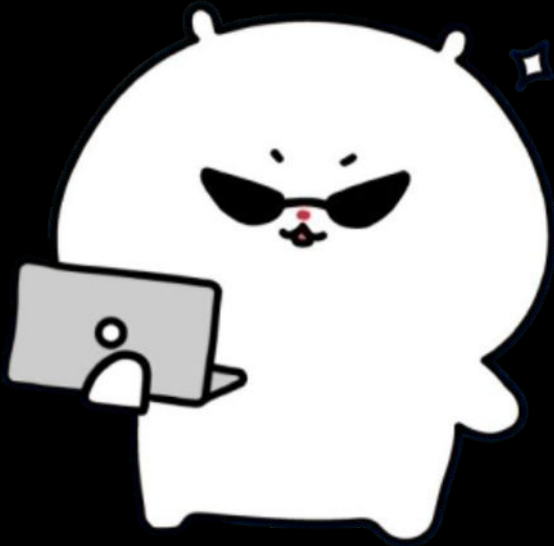
김요셉



Role

- 데이터 증강

김동규



Role

- T5실험, optuna 활용

변혜영



Role

- 데이터 증강

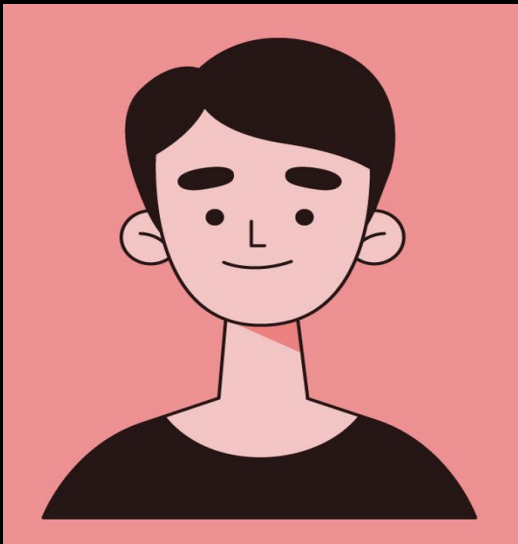
오승민



Role

- 모델 freezing 실험

이주하



Role

-back translation 및
gemma, prompt
engineering 실험

02

대회 개요

주제

Dialogue Summarization | 일상 대화 요약

개요

일상 대화는 회의, 토의, 사소한 대화를 포함해 다양한 주제와 관점을 주고받는 과정입니다.

하지만 대화를 녹음하더라도 전체를 다시 듣는 것은 비효율적이기에 요약이 필요합니다.

대화 중 요약은 집중을 방해하고, 이후 기억에 의존한 요약은 오해와 누락을 초래할 수 있습니다.

이를 해결하고자, 이번 대회에서는 일상 대화를 기반으로 자동 요약문을 생성하는 모델을 개발합니다.

대회 개요

그룹 스터디 진행 방법

각자가 실험한 내용들을 매일 오후 2시에 Zoom으로 만나서 공유하고
다음 실험의 방향성을 결정하였습니다.
앙상블을 생각하고 실험적인 방법과 안정적인 방법 모두 시도해보았습니다.

결과가 그렇게 좋지는 못했지만 다양한 방법으로 협력하는 방법을 배울 수 있었습니다:)

김동규

오승민

H

이주하

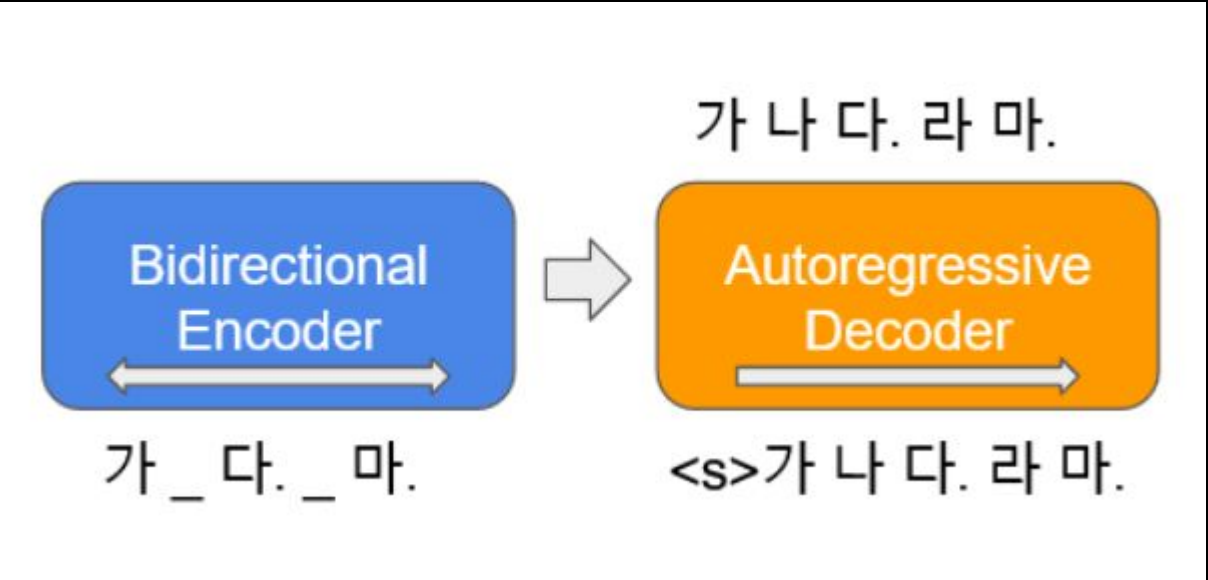
김요셉

오승민

03

모델 선정

kobart-summarization



EbanLee/kobart- summary-v3

koBART 모델은 한국어 텍스트 요약에 위한 사전 학습된 모델로 긴 한국어 텍스트를 간결하게 요약하는 데 사용

Text Infilling(문장의 빈칸 부분을 채우는 작업) 노이즈 함수를 사용하여 40GB 이상의 한국어 텍스트에 대해서 학습한 한국어 encoder-decoder 언어 모델

BART는 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 auto encoder의 형태로 학습

kobart 모델을 문서요약, 도서자료요약, 요약문 및 레포트생성 데이터로 fine-tuning 한 모델

>>결과값은 베이스보다 성능이 더 떨어짐

0.5538	0.3611	0.4691	46.1344
0.5482	0.3403	0.4509	44.6486

04

성능 최적화

성능 최적화

데이터 증강

LLM을 활용한 데이터 증강

```
# 대화 및 요약 생성 함수
def generate_dialogue_summary(topic):
    prompt = f"""
    주제: {topic}

    위 주제에 대한 대화와 요약을 생성해주세요. 대화는 #Person1#과 #Person2# 사이에서 이루어지며,
    한국어 어순 번역 말투를 사용해야 하고 요약문은 말끝이 이다 형식으로 생성한다.

    형식:
    대화:
    (여기에 대화 내용)

    요약:
    (여기에 요약 내용)
    """

    response = client.chat.completions.create(
        model="solar-pro",
        messages=[{"role": "user", "content": prompt}],
        stream=False,
        max_tokens=500
    )

    return response.choices[0].message.content
```

train.csv 파일에서 대화 topic 추출
(겹치는 topic 제외)

solar-pro 모델을 사용하여 기존 train data set에
추가적으로 topic별 대화를 생성하여 데이터 증강

```
겹치는 토픽의 개수: 215
겹치는 토픽들:
게임을 하다
아파트 임대
윈드서핑
사회적인 캐주얼 대화
호텔 예약
날씨
```

Back Translation

```
def translate_with_chatgpt_v1(texts, batch_size=10, output_file=None):
    translated_texts = []
    for i in range(0, len(texts), batch_size):
        batch = texts[i:i + batch_size]
        try:
            # Prompt 생성
            prompt = (
                "Translate the following Korean texts to English:\n" +
                "\n".join([f"{idx + 1}. {text}" for idx, text in enumerate(batch)]) +
                "\nProvide translations in the same numbered order."
            )
            # ChatGPT 호출
            response = client.chat.completions.create(
                model="gpt-4o",
                messages=[{"role": "user", "content": prompt}],
                temperature=0.3,
            )
```

- 원본 train.csv 및 dev.csv가 영어 -> 한국어로 번역된 상태여서 다시 영어로 Back translation 후에 XSum으로 영어 요약 >> 다시 한글말로 번역해보려는 시도를 했음
- 초반엔 helsinki opus mt en ko 로 작업하다 tokenizer에서 에러가 계속 나서 결국 gpt4o api로 번역품질을 높이려고 했음
- 마지막 2일동안 했던 실험이다 보니 단순 train.csv(12459개 데이터), dev.csv를 번역하는 데만 12시간 넘게 걸려서 시간 문제로 인해 마지막에 제대로 실험이 안됐음.

```
'1. ', '#Person1#: How many people are in your family?', '#Person2#: My immediate family is quite small. It's just my mother, stepfather, my biological brother, and me. How about you?', '#Person1#: I have three siblings, two brothers and one sister. They all live far away from home.', '#Person2#: That sounds like a big family. Do they visit often?', '#Person1#: Not really, mostly during holidays. We usually get together for Christmas and Thanksgiving.', '#Person2#: Sounds nice. What do you do for work?', '#Person1#: I'm a software engineer at a tech company. It's pretty busy here.', '#Person2#: Oh, cool. Do you like it?', '#Person1#: Yeah, I do. The challenges keep me on my toes.', '#Person2#: Good. Well, I've got to go now. Talk to you later!', '#Person1#: Okay, talk to you later!'
```

batch 1239: Successfully translated 10 texts.

```
'1. #Person1#: Excuse me, are you Dr. Smith?', "    #Person2#: Yes, that's correct. And you are...", "    #Person1#: I'm David, Joanna's husband. She has to work late today, so she asked me to come pick her up.", "#Person2#: No problem. I'll see you soon.", "#Person1#: Thank you very much. Have a good evening!", "#Person2#: You too. Bye-bye!"
```

batch 1240: Successfully translated 10 texts.

```
'1. ', '#Person1#: So, what are your plans for this weekend?', '#Person2#: I just want to stay at home.', '#Person1#: How about going to the movies?', "#Person2#: Sorry, I'm a bit tired these days. Maybe next time."'
```

batch 1241: Successfully translated 10 texts.

```
'1. ', "#Person1#: I watched a very interesting documentary about plants last night. It was titled 'Strange Plants' and covered plants with unique features found around the world.", '#Person2#: Really? Which ones were your favorites?', '#Person1#: The Venus flytrap was fascinating. It's almost like a predator.", "#Person2#: Wow, that sounds amazing. I'll have to watch that one too."
```

batch 1242: Successfully translated 10 texts.

```
'1. ', '#Person1#: Hello, I have an appointment with Dr. Smith that I need to reschedule.', '#Person2#: When you made the appointment, which date did you choose?', '#Person1#: My appointment was set for Friday afternoon.', '#Person2#: Okay, let me check the doctor's schedule. How about moving it to Saturday morning instead?', '#Person1#: That works perfectly. Thank you so much for your help.', '#Person2#: No worries. Feel free to call if you need anything else.'
```

batch 1243: Successfully translated 10 texts.

```
'1. ', "#Person1#: I don't know about you, but I'm so hungry I can't stand it. How about going to get something to eat?", "#Person2#: That sounds good. I'm really hungry too! What kind of food do you want to eat?", "#Person1#: I think I'll go for some pizza. There's a new place downtown that looks promising.", "#Person2#: Great choice. Let's go together."
```

batch 1244: Successfully translated 10 texts.

```
'1. #Person1#: What are you guys doing over there?', "    #Person2#: I'm not sure what you mean, Janice.", "    #Person1#: I've been waiting for a response on the Blake Building design for almost two weeks now. Any updates?", "#Person2#: Not yet, but we're working on it. Should have something by next week.", "#Person1#: Alright, thanks for letting me know. I'll follow up again then.", "#Person2#: Sure thing. See you later!"
```

batch 1245: Successfully translated 10 texts.

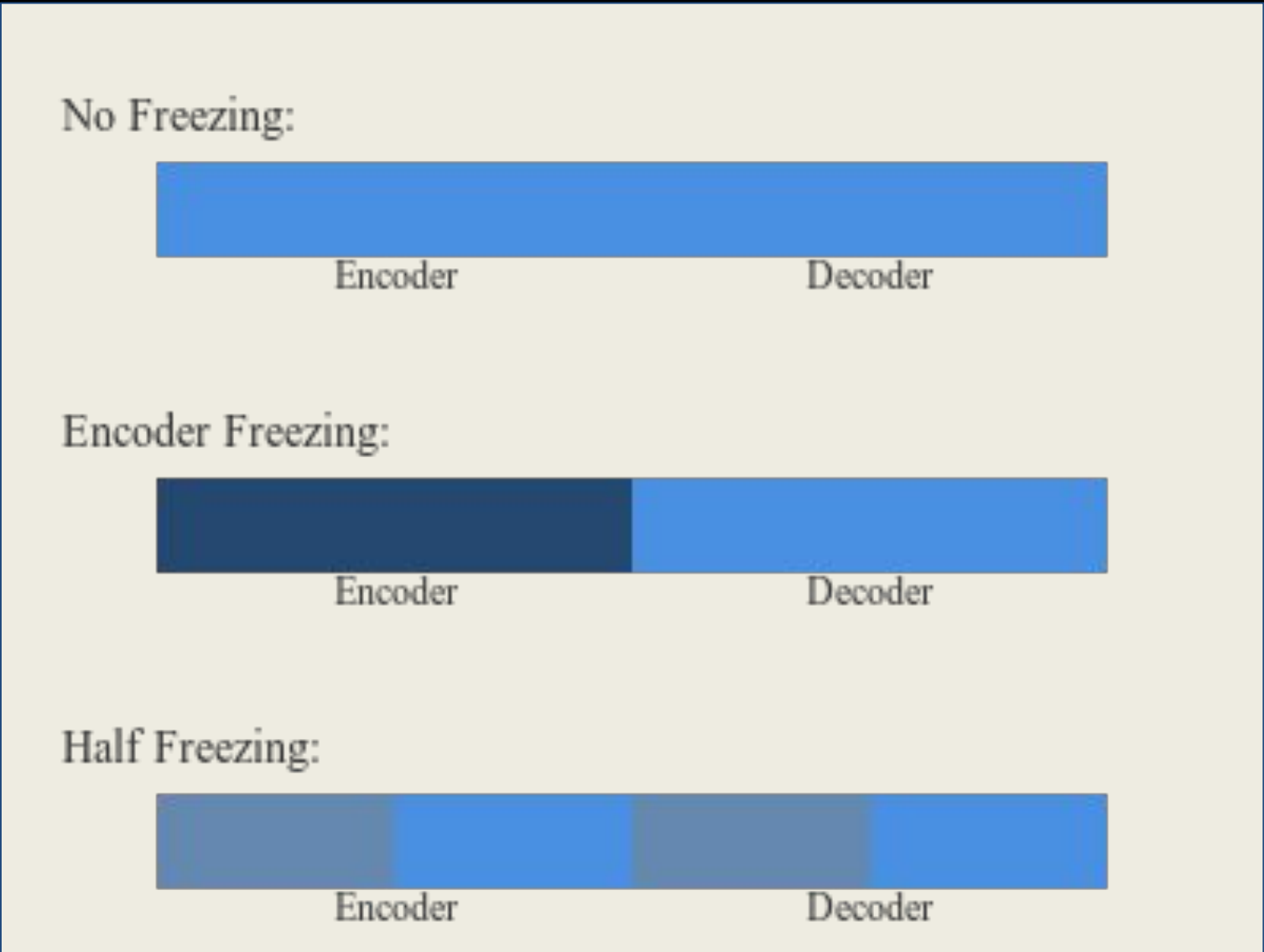
```
'1. ', '#Person1#: Dad, can I go to the movies with Sharon?', "#Person2#: Sure, but wait a minute. Weren't you supposed to get your report card this week?", '#Person1#: Um, oh, right. Can I call you back after school tomorrow?', "#Person2#: Of course. Just make sure you finish your homework first."
```

batch 1246: Successfully translated 7 texts.

```
'1. ', '#Person1#: You really look sick!', "#Person2#: I haven't been able to think clearly lately.", '#Person1#: You seem to be worried about something. Promise me, go see a doctor right away.'"
```


성능 최적화

모델 레이어 freezing



- => 모든 레이어 학습
과적합 위험 존재
- => 디코더만 학습
사전학습된 인코더 특성 보존
- => 인코더/디코더 하위 50% 레이어 고정
하위 레이어의 기본적인 특성 유지

0.5541	0.3698	0.4799	46.7944
0.5416	0.3471	0.4523	44.6997

3개 모델
== soft voting ==> 기존 koBART 보다 성능 저하
양상블 결과

05

결과 분석

결과 분석

ver. 1




baseline code에서 0.15점 향상
dev.csv에 있는 topic으로 대화문과 요약문 생성한것을 train.csv에 추가하여 학습
중간 결과값이 가장 높아 제출한 모델

<input checked="" type="checkbox"/>	gen_2750	<div>J</div>	0.5745 0.5503	0.3827 0.3456	0.4904 0.4504	48.2555 44.8750	2024.11.27 15:03	완료	<div>↓</div>
-------------------------------------	----------	--------------	------------------	------------------	------------------	--------------------	------------------	----	--------------

결과 분석

ver. 2

데이터 전처리, special tokens, 하이퍼 파라미터(학습률 등) 조정 후 적용한 kobart 모델
중간 결과값이 베이스보다 낮아서 submit 안했지만 final result가 제일 높은 결과값이 나왔음.

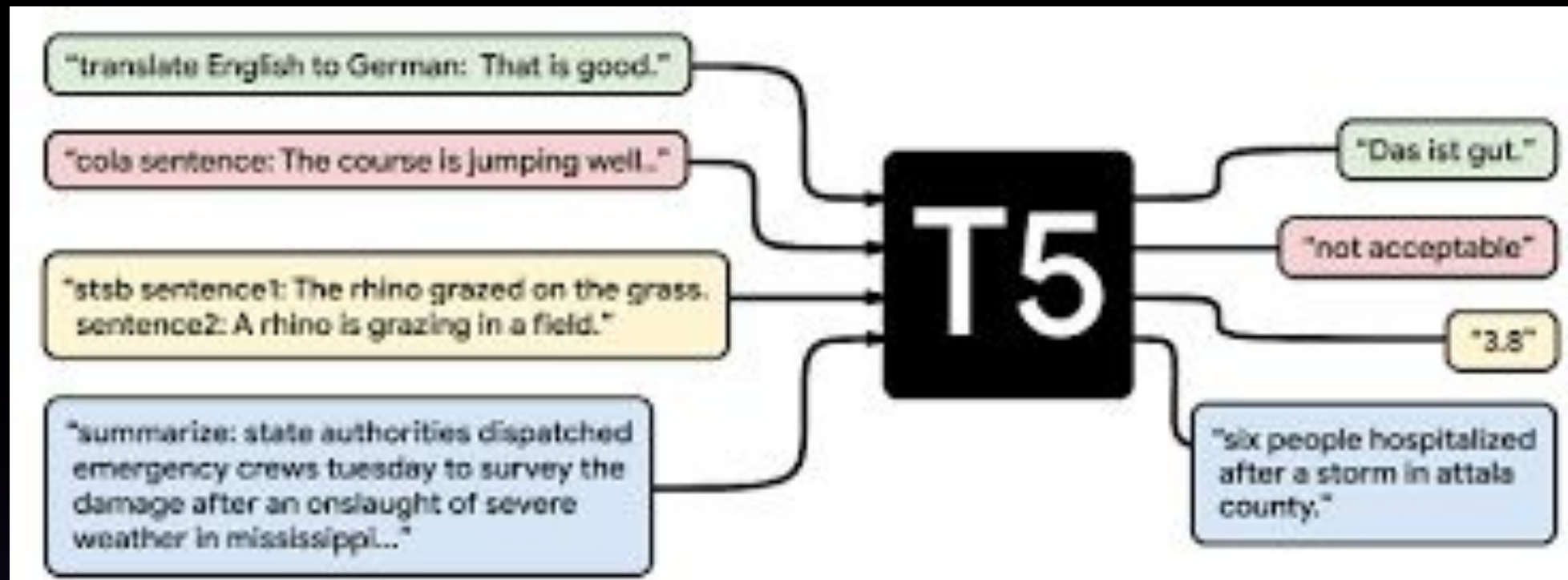
	test1		0.5590 0.5496	0.3712 0.3511	0.4736 0.4533	46.7954 45.1345	2024.11.28 17:49	완료	
---	-------	---	------------------	------------------	------------------	--------------------	------------------	----	---

06

Troubleshooting

TROUBLESHOOTING - 모델 선정

T5 Model



KoBART 말고 다른 모델로 실험하고 싶었음.
대회 사전 조사를 통해 알아보니 T5 모델로도 진행한 기록이 있었음.

keti-air/ke-t5-base 모델을 찾음
한국전자기술연구원에서 Google의 T5 모델을 활용해서 한국어와 영어가 6.5:3.5 비율로 구성된 데이터 셋을 학습시킨 모델

학습을 진행했으나 1회 학습에 많은 시간이 소요되는 문제와 의미있는 성능이 나오기 위해서는 적어도 몇번의 epoch을 돌아야 하는데 epoch을 돌다가 OOM 문제 발생.

QLora를 활용해 해결하려고 했으나 epoch 횟수는 늘어났으나 끝까지 학습을 진행하지 못하고 OOM 발생

결국... T5로 진행하는 실험은 포기...

하지만 T5의 좋은 점! prefix가 있어서 파인튜닝이 가능하다.

but 로컬 환경에서는 메모리 이슈로 prefix 길이에 제한이 있어서 원하는 만큼 파인튜닝이 어려움.

TROUBLESHOOTING - 모델 선정

Gemma 2-9b-it Model



- 다른 모델들을 검색해보다 구글 베이스의 ko-gemma 2-9b-it가 성능이 꽤 잘 나온다는 huggingface 보고 test 해보기로 결정
- 그러나 모델 크기 때문에 GPU 메모리 이슈로 자꾸 pretrain 과정에서 터졌음.
- LoRa config와 4,8bit 양자화 코드를 넣어서 시도 및 하이퍼파라미터도 많이 조정해봤지만 여전히 메모리 문제가 발생.
- 이미 bart 모델의 성능이 어느정도 검증된만큼 만약 메모리 향상을 위해 gemma 모델버전을 낮춘다고 결과가 유의미할 거 같지 않아서 실험 중단.

Model	Math	Reasoning	Writing	Coding	Understanding	Grammar	Single ALL	Multi ALL	Overall
rtzr/ko-gemma-2-9b-it	8.71 / 8.00	9.14 / 8.00	9.43 / 9.29	9.00 / 9.43	9.57 / 9.86	7.14 / 5.00	8.83	8.26	8.5
google/gemma-2-9b-it	8.57 / 7.71	8.86 / 7.00	9.29 / 9.29	9.29 / 9.57	8.57 / 8.29	6.86 / 3.86	8.57	7.62	8.0
MLP-KTlim/llama-3-Korean-Blossom-8B	6.43 / 5.71	6.86 / 5.14	9.14 / 8.57	8.29 / 8.14	8.43 / 9.29	5.71 / 5.29	7.48	7.02	7.0
yanolja/EEVE-Korean-instruct-10.8B-v1.0	5.57 / 4.29	8.14 / 5.14	8.29 / 6.29	6.43 / 7.86	9.29 / 8.57	6.57 / 3.71	7.38	5.98	6.0
allganize/Llama-3-Alpha-Ko-8B-Instruct	4.57 / 3.00	6.86 / 6.43	7.43 / 6.71	8.43 / 8.43	7.71 / 8.71	6.71 / 4.43	6.95	6.29	6.0

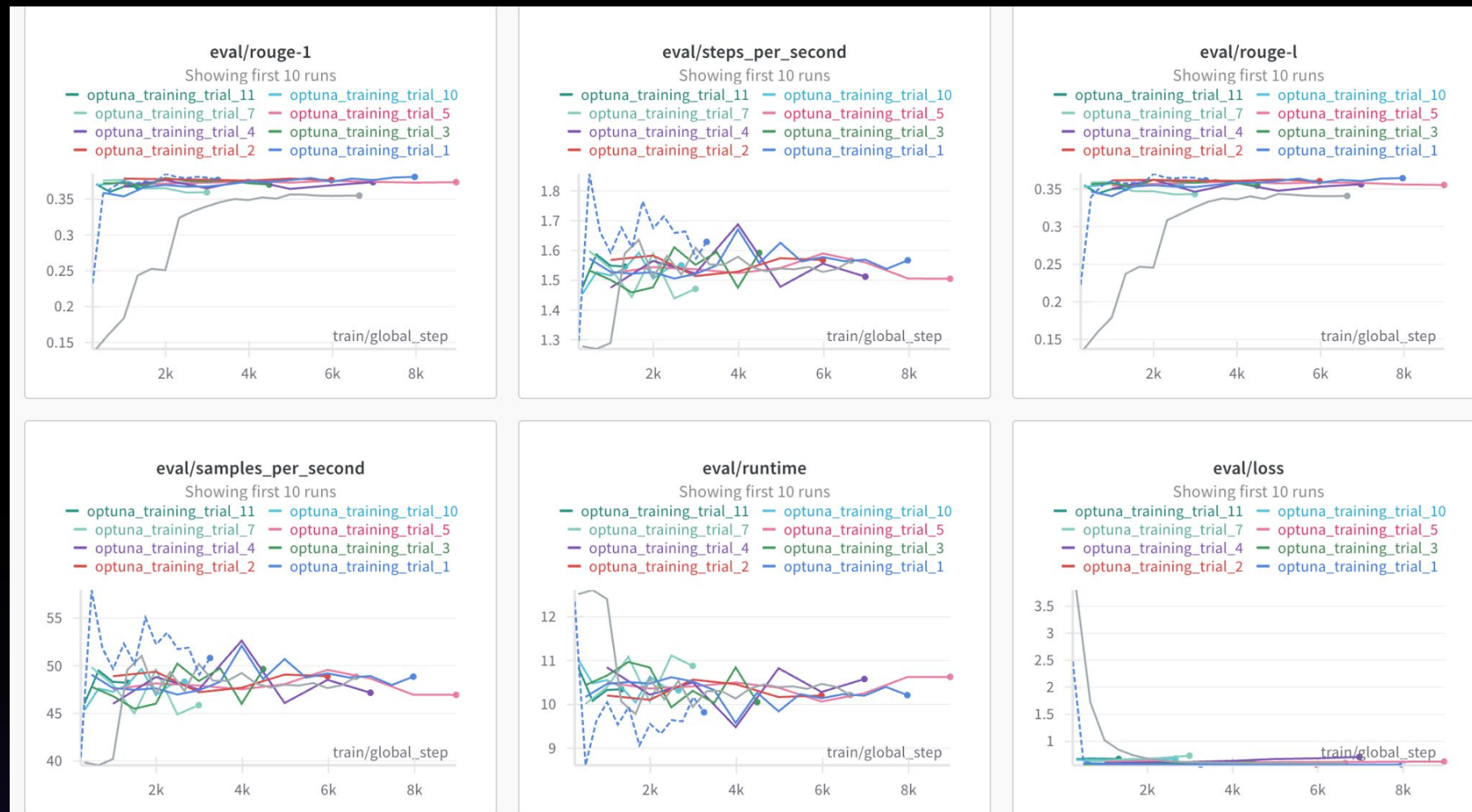
```
# LoRA Config 설정
def apply_lora(model):
    lora_config = LoraConfig(
        task_type=TaskType.SEQ_2_SEQ_LM,
        inference_mode=False,
        r=2,
        lora_alpha=8,
        lora_dropout=0.3,
        target_modules=["q_proj", "v_proj"]
    )
    lora_model = get_peft_model(model, lora_config)
    lora_model.gradient_checkpointing_enable()
    return lora_model

# 8-bit 양자화 활성화
bnb_config = BitsAndBytesConfig(
    load_in_8bit=True,
    bnb_8bit_use_double_threshold=False,
    bnb_8bit_quant_type="nf4"
)
```

TROUBLESHOOTING - 성능 최적화

Hyperparameter Optimization

Optuna



0.2724	0.1035	0.2204	19.8767
0.2724	0.0997	0.2140	19.5378

결국... overfitting
발생...

결론 : 잘못된 실험 방법이었다...

-> optuna를 활용해서 최적 파라미터를 찾을 때 여러 trial을 시도함.
하지만, 이전에 학습한 모델을 그대로 불러와서 실험하지는 않음...

그렇다면 왜 이전에 학습한 모델을 그대로 불러와서 실험했는가?

-> 데이터 증강된 train.csv파일로 실험을 했는데 데이터가 증강되면
학습 횟수도 늘어나야 하지 않을까? 생각함.

그래서 이상한 실험을 설계하고 실험함...

+ epoch에 대한 메모리 오해 이슈...

정석적인 실험 방법대로 먼저 하는지 점검하기

다음부터는 급해도 각 코드가 메모리에 미치는 영향을 바르게 알아보고 실험을 시도할 것.

06

진행 소감

그룹스터디 진행 소감

Natural Language Processing



이주하 여태 모든 대회 및 프로젝트에서 가장 메모리 문제가 가장 심했던 거 같음. 그리고 워낙 베이스라인 자체가 성능이 잘 나오다 보니 웬만한 실험을 해도 올리기 힘들었음. 영어였으면 훨씬 나왔을텐데 한국말이 아 다르고 어 다르단 걸 다시 한번 느꼈음.

김동규 이번 대회는 상당히 어려웠습니다. 메모리 부족 문제로 원하는 모델로 실험하기도 어려웠고, 각 코드가 컴퓨터의 메모리에 미치는 영향도 바르게 알지 못했고, 체계적인 가설 세우기와 실험관리를 할 수 있는 능력이 없다보니 모든 가설과 실험마다 실패를 반복했습니다. 그리고 그 어떤 대회보다 실험 실패시 시간 비용이 많이 들어갔습니다.

이번 대회를 통해 부족한 부분을 발견할 수 있어서 좋았고, 앞으로 발전할 날들이 기대가 됩니다.)

김요셉 베이스라인 코드가 다른 대회보다 점수가 높고 NLP 대회여서 그런지 이전 대회보다 점수를 향상시키기 훨씬 어려웠습니다. 파라미터를 조금만 수정해도 메모리가 부족한 문제는 다양한 실험을 하는데 발목을 잡았습니다. 팀 내에서 다양한 방법을 수행해보았지만 유의미한 결과가 없어 방향성에 대해 많이 고민하는 대회였습니다.

오승민 생각했던 성능 개선 방안들이 이번 프로젝트에서 맞지 않아서 그런건지 생각보다 성능 개선이 이루어 지지 않아서 많이 아쉬웠습니다.

변혜영 NLP는 어렵네요..데이터 증강에도 시간이 오래 걸리고..

Life-Changing Education

감사합니다.
