

Upstage AI Lab

Dialogue Summarization Competition Report

[Lucky 4th]

김정현, 임승환, 지수영

목차

1. 그룹 소개
2. 대회 소개
3. Exploratory Data Analysis
4. 모델 선정 및 경량화
5. Prompt Engineering
6. Data Augmentation
7. 결과 및 회고

01

그룹 소개



김정현

#4조의 일원 #INFJ

- prompt tuning
- Augmentation



지수영

#최선을 다하자

- 모델별 학습&실험
- 프롬프트 엔지니어링



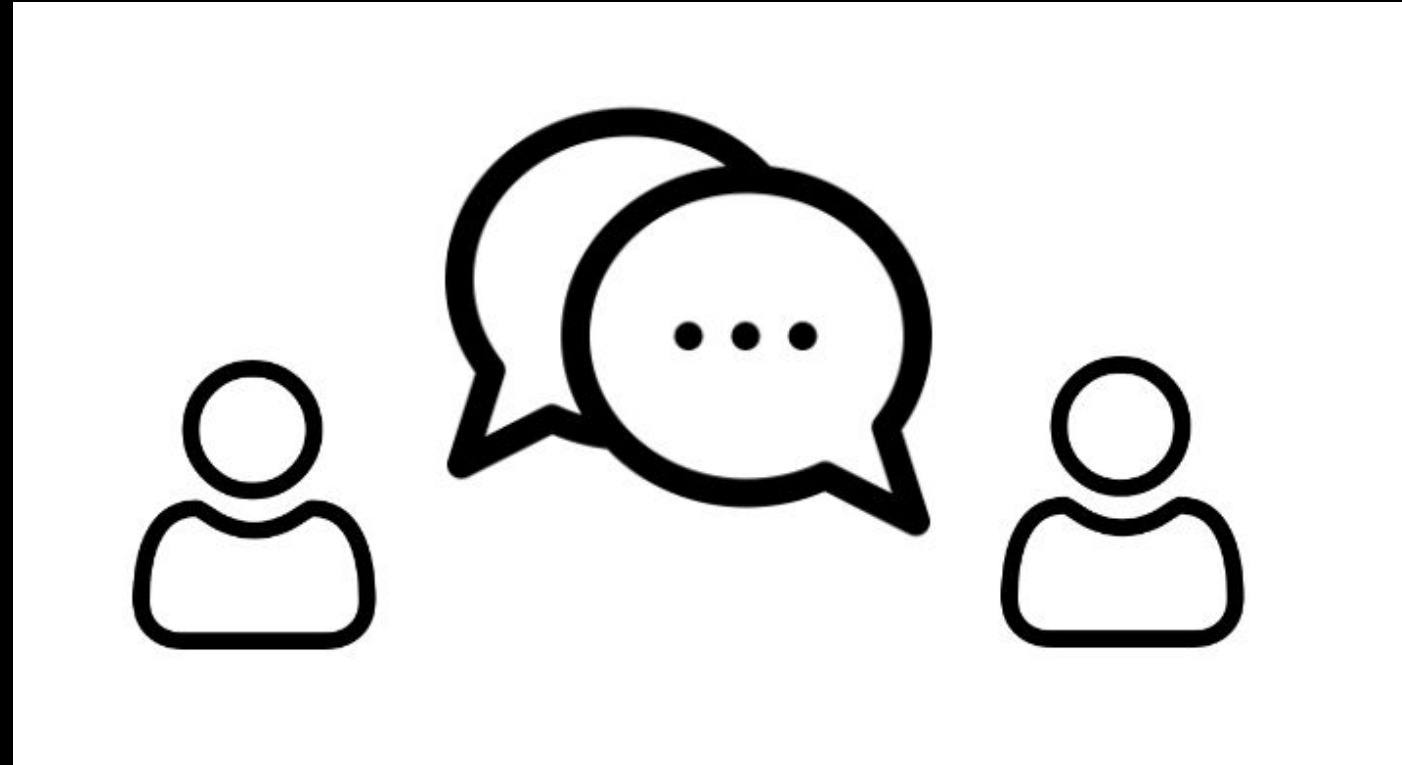
임승환

#일단한다

- SOLAR base 코드 작성
- 모델 학습
- 하이퍼 파라미터 튜닝
- BackTranslate
- Topic generation

02

대회 소개



“일상 대화 요약 모델 개발 경진대회”

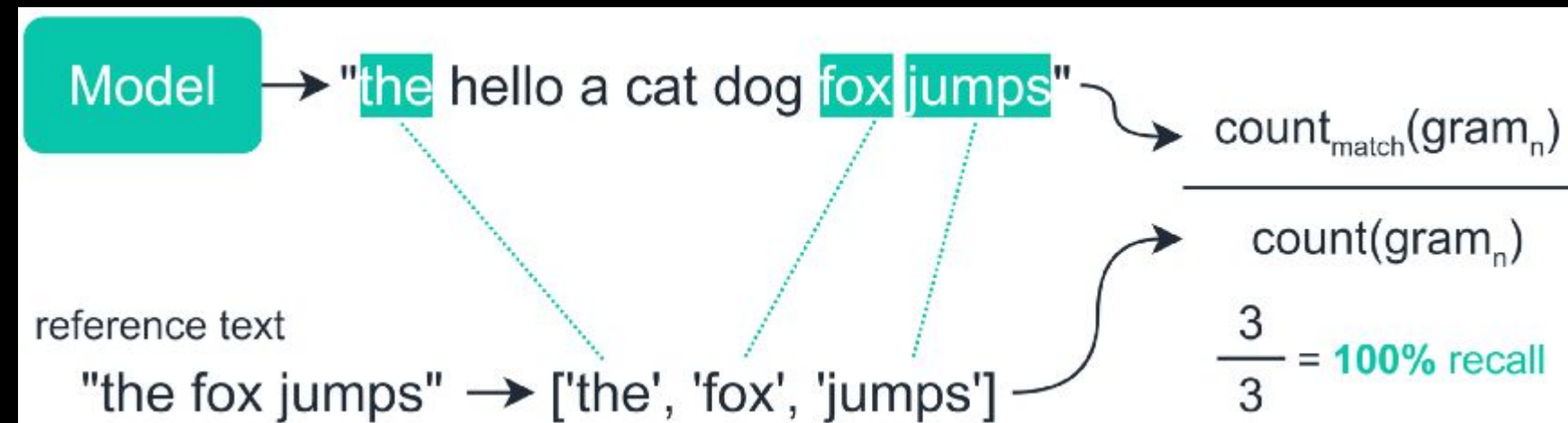
- 일상 대화를 바탕으로 요약문 생성하는 모델 구축
- 다양한 구조의 자연어 모델 활용 가능
- 정확하고 일반화된 요약 모델 개발

	fname	dialogue	summary	topic
0	train_0	#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나...	스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니...	건강검진 받기
1	train_1	#Person1#: 안녕하세요, 파커 부인, 어떻게 지내셨나요?\n#Person2#...	파커 부인이 리키를 데리고 백신 접종을 하러 갔다. 피터스 박사는 기록을 확인한 후...	백신
2	train_2	#Person1#: 실례합니다, 열쇠 한 묶음 보셨나요?\n#Person2#: 어떤...	#Person1#은 열쇠 한 묶음을 찾고 있고, 그것을 찾기 위해 #Person2#...	열쇠 찾기
3	train_3	#Person1#: 왜 너는 여자친구가 있다는 걸 말해주지 않았어?\n#Person...	#Person1#은 #Person2#가 여자친구가 있고 그녀와 결혼할 것이라는 사실...	여자친구가 있다
4	train_4	#Person1#: 안녕, 숙녀분들! 오늘 밤 당신들은 정말 멋져 보여. 이 춤을 ...	말릭이 니키에게 춤을 요청한다. 말릭이 발을 밟는 것을 신경 쓰지 않는다면 니키는 ...	댄스
...
12452	train_12455	#Person1#: 실례합니다. 맨체스터 출신의 그린 씨이신가요?\n#Person2...	탄 링은 흰머리와 수염으로 쉽게 인식되는 그린 씨를 만나 호텔로 데려갈 예정입니다....	누군가를 태우다
12453	train_12456	#Person1#: 이윅 씨가 우리가 컨퍼런스 센터에 오후 4시에 도착해야 한다고 ...	#Person1#과 #Person2#는 이윅 씨가 늦지 않도록 요청했기 때문에 컨퍼...	컨퍼런스 센터
12454	train_12457	#Person1#: 오늘 어떻게 도와드릴까요?\n#Person2#: 차를 빌리고 싶...	#Person2#는 #Person1#의 도움으로 5일 동안 소형 차를 빌립니다.	차 렌트
12455	train_12458	#Person1#: 오늘 좀 행복해 보이지 않아. 무슨 일 있어?\n#Person2...	#Person2#의 엄마가 일자리를 잃었다. #Person2#는 엄마가 우울해하지 ...	실직
12456	train_12459	#Person1#: 엄마, 다음 토요일에 이 삼촌네 가족을 방문하기 위해 비행기를 ...	#Person1#은 다음 토요일에 이 삼촌네를 방문할 때 가방을 어떻게 싸야 할지 ...	짐 싸기

[데이터 건수]

- train : 12457
- dev : 499
- test : 250
- hidden-test : 249

- name : 대화 고유번호
- dialogue : 최소 2명에서 최대 7명이 등장하는 대화
- summary : 해당 대화를 바탕으로 작성된 요약문
- topic : 해당 대화의 주제

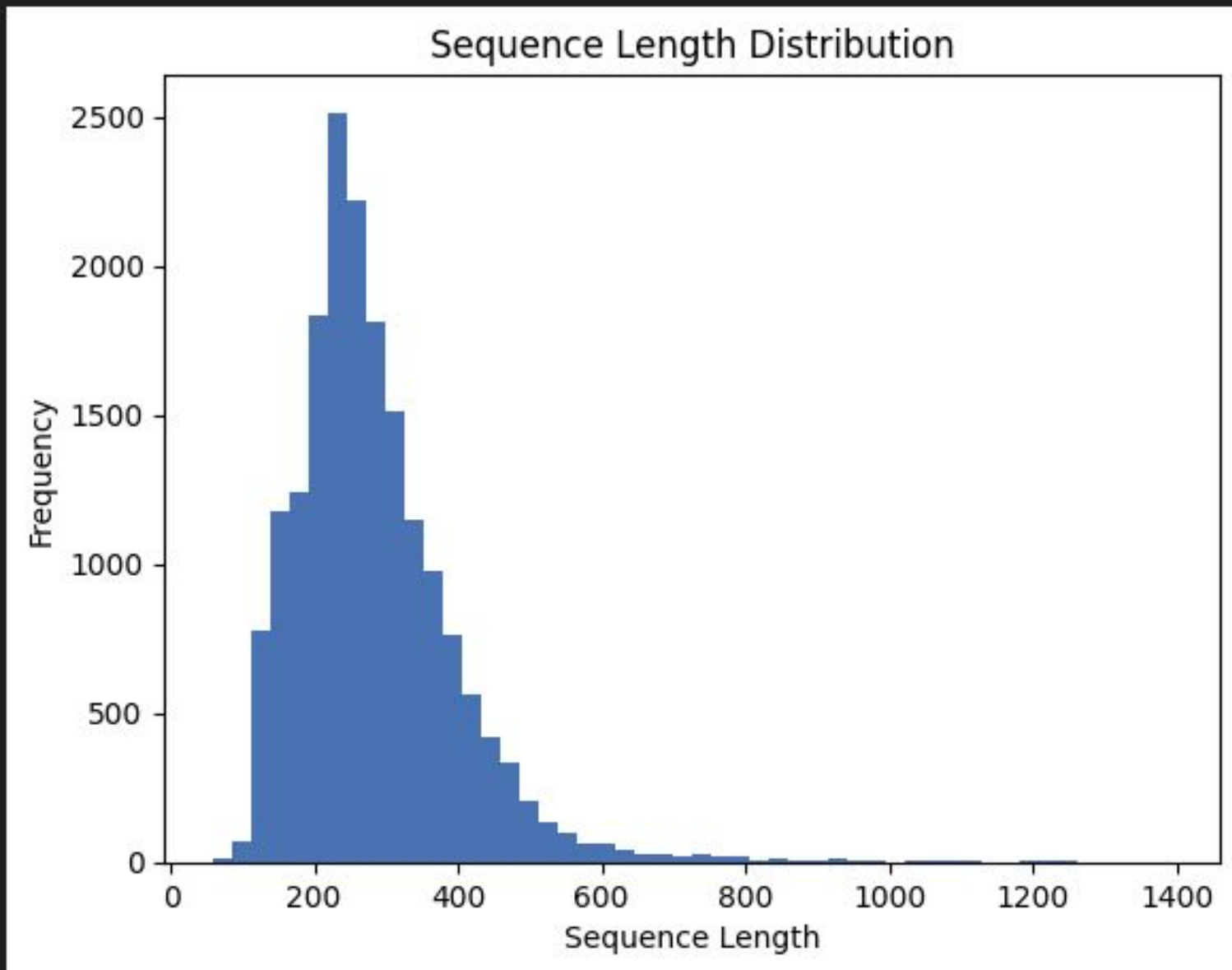


- 본 대회에서는 ROUGE-1-F1, ROUGE-2-F1, ROUGE-L-F1, 총 3가지 종류의 metric으로부터 산출된 평균 점수를 더하여 최종 점수를 계산
- ROUGE는 텍스트 요약, 기계 번역과 같은 태스크를 평가하기 위해 사용되는 대표적인 metric
모델 요약본 혹은 번역본을 사람이 만든 참조 요약본과 비교해 점수를 계산

03

Exploratory Data Analysis

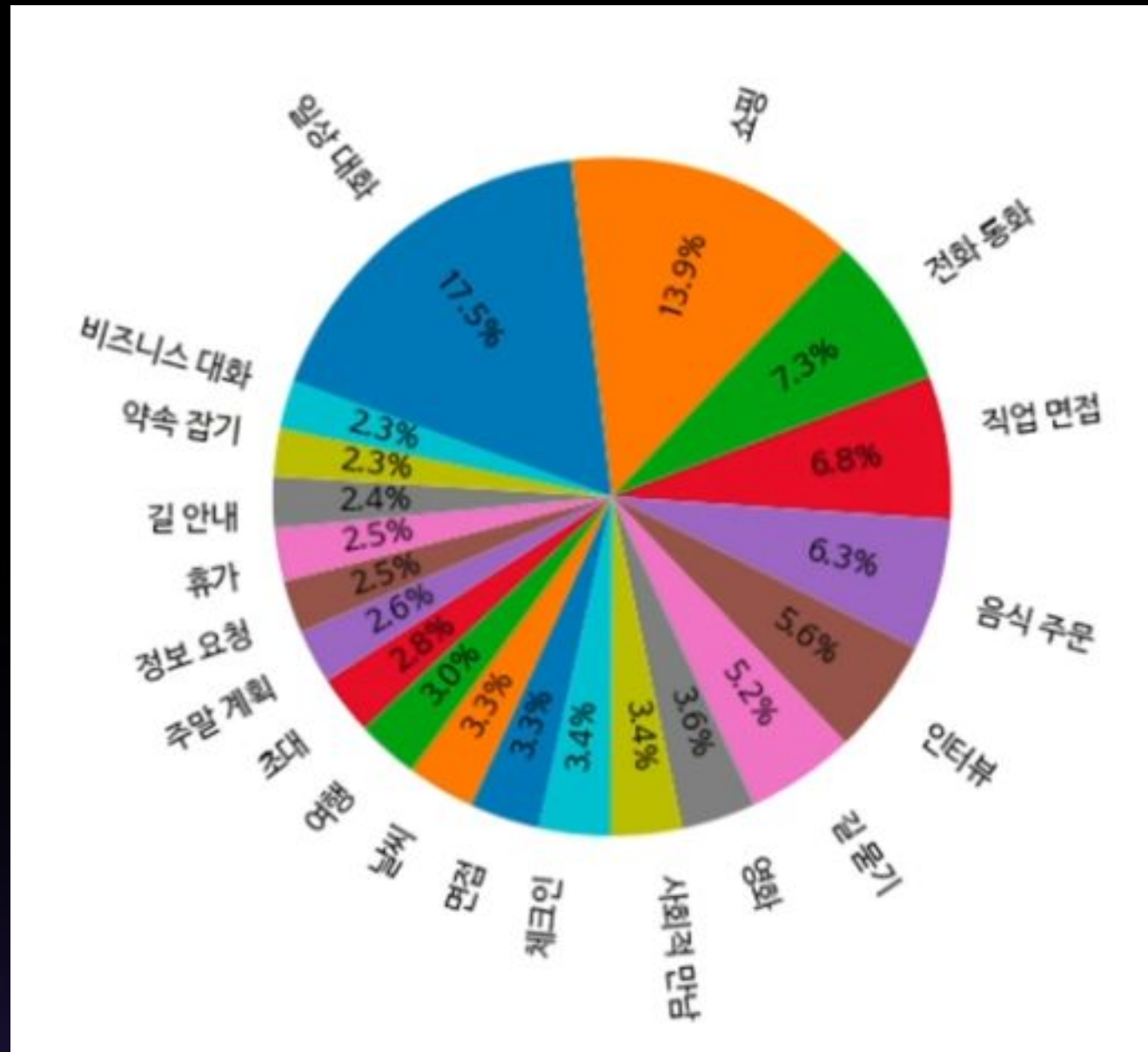
Exploratory Data Analysis



- 토큰라이저를 사용해 프롬프트 시퀀스 길이를 확인
 - 평균 시퀀스 길이: 283.8
 - 중앙값 시퀀스 길이: 263.0
 - 최대 시퀀스 길이: 1395
 - 최소 시퀀스 길이: 59

Exploratory Data Analysis

❖ 주요 대화 주제



- 대부분의 대화가 일상에서 일어나는 주제
- 비즈니스 대화, 직업 면접 등 비즈니스 상황에서 일어나는 대화도 존재

04

모델 선정 및 경량화

LLM - chihoonlee10/T3Q-ko-solar-dpo-v7.0

- Decoder model
- Open ko-LLM LB#1
- Open LB Score : 40.3

LLM - beomi/OPEN-SOLAR-KO-10.7B

- Decoder model
- Open LB Score : 44.6

LLM - beomi/gemma-ko-7b

- Decoder model
- Open LB Score : 43.4

LLM - beomi/Solar-Ko-Recovery-11B

- Decoder model
- Open LB Score : 44.9



Solar-Ko-Recovery-11B 최종 선정

- 가장 큰 모델 사이즈
- 가장 높은 점수

경량화 - PEFT (Parameter-Efficient Fine-Tuning)

- 대규모 언어 모델의 파인튜닝을 효율적으로 수행하기 위한 기법
- 기존 모델의 대부분의 가중치를 고정하고, 소량의 추가 파라미터만을 도입하여 파인튜닝
- 계산 효율과 메모리 효율을 높이는 것이 목적
- 대표적인 PEFT 기법: Adapter, Prefix-tuning, LoRA 등

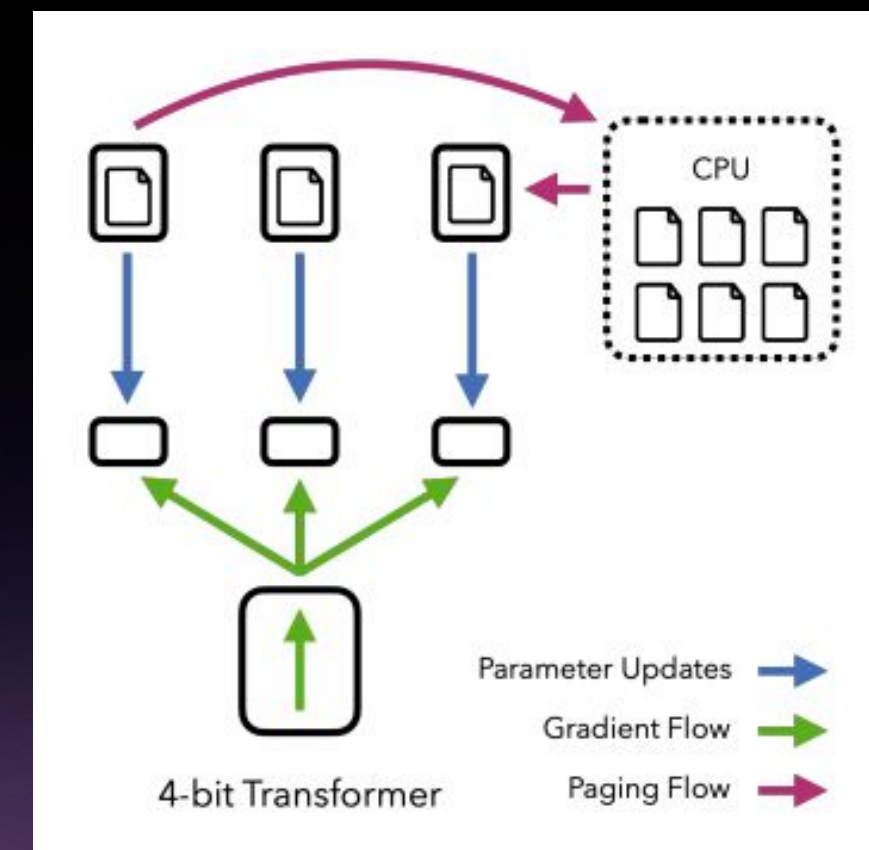


경량화 - QLoRA (Quantized Low-Rank Adaptation)

- PEFT 기법 중 하나로, LoRA에 양자화 기법을 적용한 방법
- Fine-tuning 단계에서 적은 수의 추가 파라미터를 도입, 효과적인 모델 적응을 목적
- 기존 모델의 가중치는 고정, 추가된 Low-Rank 파라미터만 업데이트해 계산 효율 향상
- 모델 크기 증가를 최소화하면서도 태스크 특화된 성능 향상을 이끌어냄

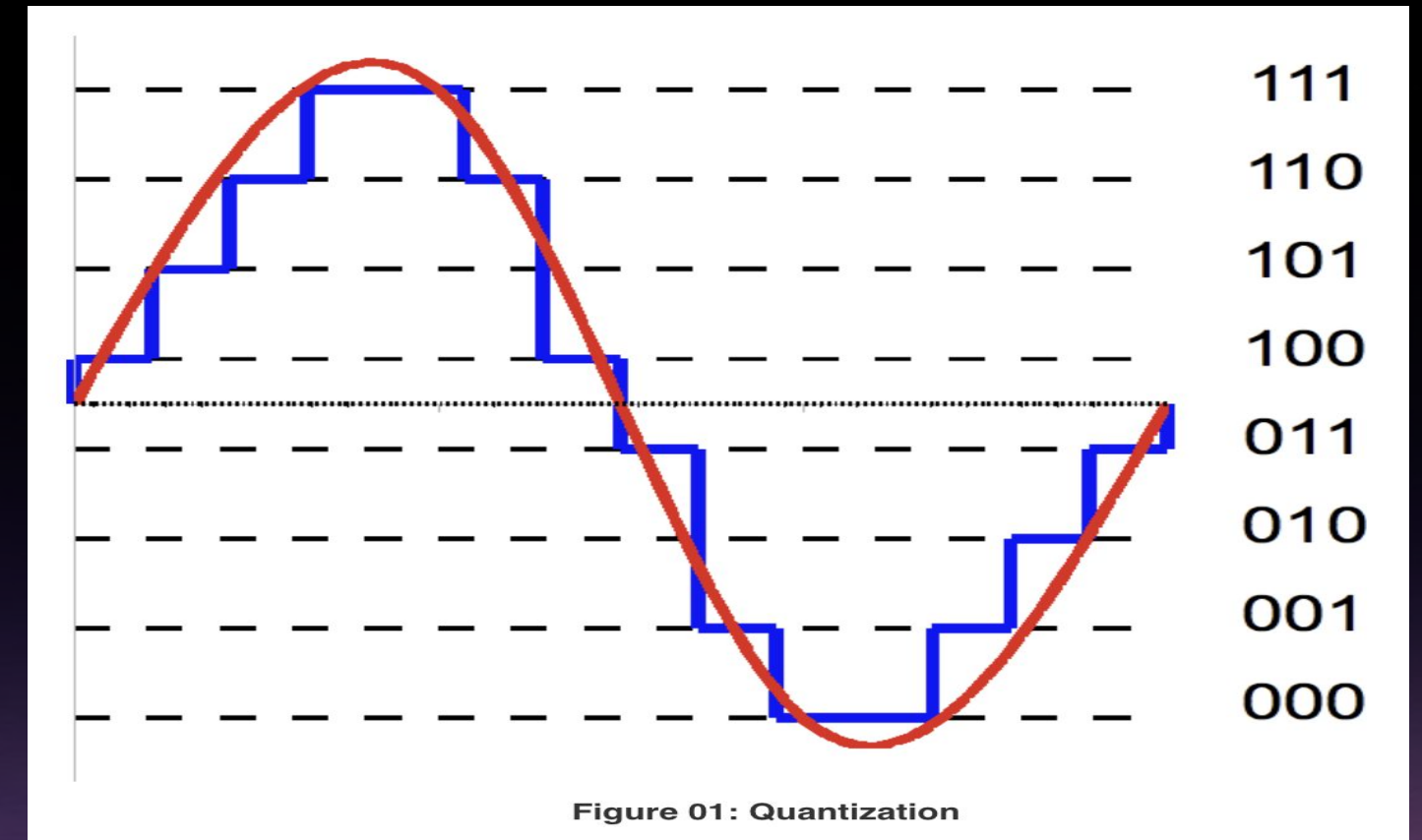
LoRA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

Edward Hu* Yelong Shen* Phillip Wallis Zeyuan Allen-Zhu
Yuanzhi Li Shean Wang Lu Wang Weizhu Chen
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)



경량화 - 4bit 양자화 (Quantization)

- 모델 가중치의 숫자 표현을 줄여 메모리 사용량을 감소시키는 기법
- 4bit QLoRA에서는 4-bit NormalFloat Quantization(NF4)를 사용
- NF4는 1bit 부호, 2bit 지수, 1bit 유효 숫자로 구성된 4bit 양자화 방법
- NF4를 통해 가중치를 효과적으로 양자화하여 메모리 사용량을 약 1/8로 감소
- NF4는 하드웨어 친화적인 설계로 효율적인 연산 가능



경량화 - 4bit 양자화 (Quantization)

- 모델 가중치의 숫자 표현을 줄여 메모리 사용량을 감소시키는 기법
- 4bit QLoRA에서는 4-bit NormalFloat Quantization(NF4)를 사용
- NF4는 1bit 부호, 2bit 지수, 1bit 유효 숫자로 구성된 4bit 양자화 방법
- NF4를 통해 가중치를 효과적으로 양자화하여 메모리 사용량을 약 1/8로 감소
- NF4는 하드웨어 친화적인 설계로 효율적인 연산 가능

05

Prompt Engineering

Prompt Engineering

```
def generate_prompt(row):
    topic = row['topic'] if 'topic' in row else ''

    prompt = f"""<s> <|im_start|>system\n주제를 참고해 다음 대화를 요약해주세요
:\n\ntopic: {topic}\n\ndialogue: {row['dialogue']}\n<|im_end|>

\n<|im_start|>assistant\n요약: {row['summary']}\n<|im_end|> </s>"""

    return prompt
```

기존 프롬프트

44.9821	0.5383	0.3549	0.4563
---------	--------	--------	--------

```
def generate_prompt(row):
    topic = row['topic'] if 'topic' in row else ''

    prompt = f"""<s> <|im_start|>system\ntopic을 참고해 다음 dialogue를 요약해주세요.아래의 사항을 지켜야 합니다.

    1.관찰자의 관점에서 작성
    2.대화의 가장 중요한 정보를 전달
    3.대화길이의 20% 이내로 요약
    4.대화 내에서 중요하게 명명된 개체를 보존
    5.문어나 약어 없이 공식적으로 사용하는 언어로 작성

:\n\ntopic: {topic}\n\ndialogue: {row['dialogue']}\n<|im_end|>\n<|im_start|>assistant\n요약: {row['summary']}\n<|im_end|> </s>"""

    return prompt
```

지시 사항 추가

44.8116	0.5359	0.3523	0.4562
---------	--------	--------	--------

```
def generate_prompt(row,example_input,example_output):
    topic = row['topic'] if 'topic' in row else ''

    prompt = f"""<s><|im_start|>system\nExample과 topic을 참고해 다음 dialogue를 요약해주세요. 아래의 사항을 지켜야 합니다.

    1.관찰자의 관점에서 작성
    2.대화의 가장 중요한 정보를 전달
    3.대화길이의 20% 이내로 요약
    4.대화 내에서 중요하게 명명된 개체를 보존
    5.문어나 약어 없이 공식적으로 사용하는 언어로 작성

Example:
    Input: {example_input}
    Output: {example_output}

:\n\ntopic: {topic}\n\ndialogue: {row['dialogue']}\n<|im_end|>\n<|im_start|>assistant\n요약: {row['summary']}\n<|im_end|> </s>"""

    return prompt
```

지시사항+예시

40.4947	0.4856	0.3108	0.4184
---------	--------	--------	--------

Prompt Engineering

- 기존 추론 시 prompt

"<s><|im_start|>system

다음 대화를 요약해주세요:

dialogue: 대화문

<|im_end|>

<|im_start|>assistant

요약:"

45.7659 0.5466 0.3609 0.4655

- topic 추가 prompt

<s><|im_start|>system

주제를 참고해 다음 대화를

요약해주세요:

topic: 주제

dialogue: 대화문

<|im_end|>

<|im_start|>assistant

요약:"

45.8550 0.5463 0.3622 0.4672

test data의 topic을 찾도록 solar를 fine-tuning
찾은 topic을 test data에 합쳐서 prompt 생성

06

Data Augmentation

Data Augmentation

EDA(Easy Data Augmentation)

- **SR(Synonym Replacement):** 문장에서 불용어를 제외한 임의의 단어를 선택 후, 동의어로 대체
- **RI(Random Insertion):** 문장의 임의의 위치에 임의의 단어를 삽입
- **RS(Random Swap):** 문장의 임의의 두 단어의 위치를 스왑
- **RD(Random Deletion):** 문장내의 임의의 단어를 삭제

'#Person1#: 안녕하세요, 실례지만 무엇을 도와드릴까요?\\n#Person2#: 저는 908호의 벨입니다. 방을 바꿔주실 수 있나요? 아내가 밤새 끔찍한 소음 때문에 여러 번 깨어났는데, 그녀가 너무 힘들어하네요.\\n#Person1#: 정말 죄송합니다. 908호는 복도 끝에 위치해 있어서 아침 일찍 소음이 들릴 수 있습니다.\\n#Person2#: 어쨌든, 방을 바꾸고 싶습니다.\\n#Person1#: 문제 없습니다, 저희가 처리하겠습니다.하지만 오늘은 모든 방이 예약되어 있어서, 내일까지 기다려주실 수 있나요?\\n#Person2#: 알겠습니다, 내일 저녁에는 조용한 방에서 편안하게 머무르고, 숙면을 취할 수 있기를 바랍니다.\\n#Person1#: 네, 그렇게 처리하겠습니다. 더 필요한 것이 있다면 알려주세요.'

#Person1#: 안녕하세요, 실례지만 무엇을 도와드릴까요? #Person2#: 저는 908호의 벨입니다. 방을 바꿔주실 수 있나요? 아내가 밤새 끔찍한 때문에 여러 번 깨어났는데, 그녀가 힘들어하네요. #Person1#: 정말 죄송합니다 요행. 908호는 복도 끝 수익에 위치해 있어서 아침 일찍 소음이 들릴 수 있습니다. #Person2#: 어쨌든, 방 바꾸고 싶습니다. #Person1#: , 없습니다 저희가 처리하겠습니다.하지만 오늘은 모든 방이 예약되어 있어서, 내일까지 기다려주실문제수 있나요? #Person2#: 알겠습니다, 내일 저녁에는 조용한 방에서 편안하게 머무르고, 숙면을있기취할 수 바랍니다. #Person1#: 네, 그렇게 처리하겠습니다. 더 필요한 것 있다면 .

Data Augmentation

AEDA(An Easier Data Augmentation)

- **[0, 전체 단어의 갯수의 1/3]** 내에서 무작위로 문장부호(.?:!.,;)를 넣어서 데이터 증강

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요? #Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다. #Person2#: 알고 있습니다. 하지만 아무 문제가 없다면 왜 의사를 만나러 가야 하나요? #Person1#: 심각한 질병을 피하는 가장 좋은 방법은 이를 조기에 발견하는 것입니다. 그러니 당신의 건강을 위해 최소한 매년 한 번은 오세요. #Person2#: 알겠습니다. #Person1#: 여기 보세요. 당신의 눈과 귀는 괜찮아 보입니다. 깊게 숨을 들이쉬세요. 스미스씨, 담배 피우시나요? #Person2#: 네. #Person1#: 당신도 알다시피, 담배는 폐암과 심장병의 주요 원인입니다. 정말로 끊으셔야 합니다.

#Person1#: 안녕하세요, 스미스씨: . 저는 , 호킨스 의사 ! 입니다. 오늘 왜 오셨나요? #Person2#: : 건강검진 ; 을 받는 . 것이 좋을 것 같아서요. #Person1#: ! 그렇군요, . 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 ! 받아야 ? 합니다. #Person2#: 알고 있습니다. . 하지만 아무 문제가 ! 없다면 왜 의사를 만나러 가야 ! 하나요? , #Person1#: 심각한 질병을 피하는 가장 좋은 방법은 이를 조기에 발견하는 것입니다. 그러니 당신의 건강을 위해 최소한 매년 한 번은 , 오세요 ; . #Person2#: : 알겠습니다. #Person1#: ; 여기 보세요 . . 당신의 눈과 귀는 괜찮아 보입니다. 깊게 숨을 들이쉬세요. 스미스씨, 담배 피우시나요? #Person2#: 네. ? #Person1#: 당신도 알다시피, 담배는 폐암과 심장병의 주요 원인입니다 ! . 정말로 끊으셔야 합니다. ?

Data Augmentation

Back Translation

- 문장을 임의의 외국어로 번역 후 원문 언어로 재번역하여 데이터 증강
- **google translate api** (한 - 영 - 한)

"#Person1#: 안녕하세요, 파커 부인, 어떻게 지내셨나요?#Person2#: 안녕하세요, 피터스 박사님. 잘 지냈습니다, 감사합니다. 리키와 함께 백신 접종을 위해 왔습니다.#Person1#: 좋습니다. 백신 접종 기록을 보니, 리키는 이미 소아마비, 디프테리아, B형 간염 백신을 맞았군요. 그는 14개월이므로, 이제 A형 간염, 수두, 홍역 백신을 맞아야 합니다.#Person2#: 풍진과 볼거리는 어떻게 되나요?#Person1#: 지금은 이 백신들만 접종할 수 있고, 몇 주 후에 나머지를 접종할 수 있습니다.#Person2#: 좋습니다. 박사님, 저도 디프테리아 예방접종이 필요할 것 같아요. 마지막으로 맞은 게 아마도 15년 전이었던 것 같아요!#Person1#: 저희가 기록을 확인하고 간호사에게 부스터를 접종하도록 하겠습니다. 이제, 리키의 팔을 꼭 잡아주세요, 조금 찌릿할 수 있습니다."

"#Person1#: 안녕하세요, 파커 부인, 잘 지내세요?#Person2#: 안녕하세요, 피터스 박사. 감사합니다. 저는 예방 접종을 위해 Ricky와 함께 왔습니다.#Person1#: 예방 접종 기록에 따르면 Ricky는 이미 소아마비, 디프테리아 및 B 형 간염 백신을 쳤습니다. 14 개월이므로 A 간염, 수두 및 홍역 백신이어야합니다.#Person2#: 풍진과 명소는 무엇입니까?#Person1#: 이제 이러한 백신 만 접종 할 수 있으며 몇 주 후에 나머지는 접종 할 수 있습니다.#Person2#: Great.dr., 나는 디프테리아 예방 접종이 필요하다고 생각합니다. 내가 마지막으로 타격을 입은 것은 아마도 15 년 전이었습니다!#Person1#: 우리는 기록을 확인하고 간호사의 부스터를 접종 할 것입니다."

Data Augmentation

Back Translation

- 문장을 임의의 외국어로 번역 후 원문 언어로 재번역하여 데이터 증강
- **papago web crawling** (한 - 일 - 한)

#Person1#: 안녕하세요, 파커 부인, 어떻게 지내셨나요?
#Person2#: 안녕하세요, 피터스 박사님. 잘 지냈습니다, 감사합니다. 리키와 함께 백신 접종을 위해 왔습니다.
#Person1#: 좋습니다. 백신 접종 기록을 보니, 리키는 이미 소아마비, 디프테리아, B형 간염 백신을 맞았군요. 그는 14개월이므로, 이제 A형 간염, 수두, 홍역 백신을 맞아야 합니다.
#Person2#: 풍진과 볼거리는 어떻게 되나요?
#Person1#: 지금은 이 백신들만 접종할 수 있고, 몇 주 후에 나머지를 접종할 수 있습니다.
#Person2#: 좋습니다. 박사님, 저도 디프테리아 예방접종이 필요할 것 같아요. 마지막으로 맞은 게 아마도 15년 전이었던 것 같아요!
#Person1#: 저희가 기록을 확인하고 간호사에게 부스터를 접종하도록 하겠습니다. 이제, 리키의 팔을 꼭 잡아주세요, 조금 찌릿할 수 있습니다.

#Person1#: 안녕하세요, 파커 부인, 어떻게 지내셨어요?
#Person2#: 안녕하세요, 피터스 박사님. 잘 지냈어요, 감사합니다. 리키와 함께 백신 접종을 위해 왔습니다.
#Person1#: 좋습니다. 백신 접종 기록을 보면 리키는 이미 소아마비, 디프테리아, B형 간염 백신을 맞고 있군요. 그는 14개월이므로 앞으로는 A형 간염, 수두, 홍역 백신을 맞아야 합니다.
#Person2#: 풍진과 볼거리는 무엇입니까?
#Person1#: 지금은 이 백신들만 접종할 수 있고, 몇 주 후에 나머지를 접종할 수 있습니다.
#Person2#: 좋습니다. 박사님, 저도 디프테리아 예방접종이 필요하다고 생각을 합니다. 마지막으로 맞은 건 아마 15년 전이었을 거예요!
#Person1#: 저희가 기록을 확인하고 간호사에게 부스터를 맞을 거예요. 이제 리키의 팔을 꼭 잡으세요, 조금 찌릿찌릿할 수 있어요.

06

결과 및 회고

Final Model config

Lora config

```
r=6,  
target_modules=["q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj"],  
task_type="CAUSAL_LM",
```

BnB config


```
load_in_4bit=True,  
bnb_4bit_quant_type="nf4",  
bnb_4bit_compute_dtype=torch.float16
```

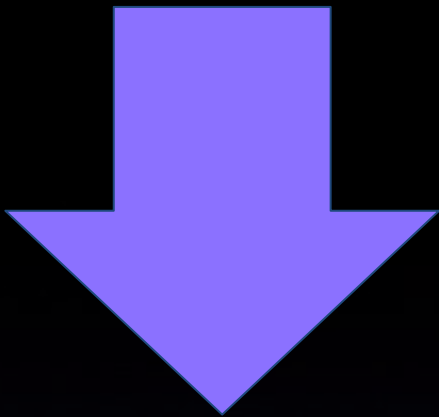
Training Argument





```
max_seq_length=512  
num_train_epochs=3,  
per_device_train_batch_size=1,  
per_device_eval_batch_size=1,  
gradient_accumulation_steps=4,  
eval_accumulation_steps=4,  
optim="adamw_torch_fused",  
warmup_steps=0.05,  
learning_rate=2e-4,  
fp16=True,  
max_grad_norm=0.3,  
weight_decay=0.001,
```

+original train data
+original validation data
+test data with topic

Final Leaderboard Score

순위	팀 이름	팀 멤버	final_result ↕	rouge1 ↕	rouge2 ↕	rougeL ↕	제출 횟수
2 (-)	NLP-04	   	45.8550	0.5463	0.3622	0.4672	29



순위	팀 이름	팀 멤버	final_result ↕	rouge1 ↕	rouge2 ↕	rougeL ↕	제출 횟수
3 (1 ▼)	NLP-04	   	42.8550	0.5217	0.3282	0.4357	29

그룹 스터디 인사이드 공유

1

인사이드 1



제한된 성능내에서의
모델선택 어려움

2

인사이드 2



LLM에는 단순한
증강데이터는 별로
의미가 없다

3

인사이드 3



프롬프트 엔지니어링이
생각보다 중요



지수영 제한된 성능에서 맞는 모델을 찾아 돌리는게 시간도 오래 걸리고 어려웠다. 그리고 기간이 짧아서 아쉬웠다.

김정헌 점수가 안올라요

임승환 LLM fine tuning을 처음 해보면서 많은공부가 되었던 것 같다. 점수가 굉장히 천천히 올라서 심리적으로 힘든 부분이 있었던 것 같다.

감사합니다. 그리고
다들 고생 많으셨습니다!