SK네트웍스 Family AI 과정 19기

데이터 전처리 인공지능 학습 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 학습 결과서
제출 일자	2026-01-12
깃허브 경로	https://github.com/orgs/Poli-Cheetah
작성 팀원	박준영

1. 모델 비교 및 선정 이유

- 비교 대상 모델:

모델명	종류	선정 이유
google/gemma-2-2b-it	Transformer 기반 사전학습 모델	동작 시간이 오래 걸리지 않는 소형 모델, 다양한 사이즈의 동일 계열 모델 존재 - 지식 증류 등의 학습 방식 사용 가능, 테스트 모델 중 가장 뛰어난 성능
Qwen/Qwen3-1.7B	Transformer 기반 사전학습 모델	
SEOKDONG/llama3.2_1B_korean_v0.2_sft_by_aidx	Transformer 기반 사전학습 모델	
upstage/SOLAR-10.7B-Instruct-v1.0	Transformer 기반 사전학습 모델	

- 최종 선정 모델: google/gemma-2-2b-it
- 모델의 선정에는 distillation을 통한 학습 가능성을 최우선으로 고려하며, 모델의 기본 성능과 사이즈를 중심적으로 고려

2. 모델 구조 및 아키텍처

2.1 모델 아키텍처 도식 (선택사항: 도식 첨부 또는 말로 설명)

Input Tokens

↓

Token Embedding

↓

[Transformer Block × N]

├─ RMSNorm

├─ Grouped Multi-Head Self-Attention (RoPE)

├─ Residual Add

├─ RMSNorm

├─ SwiGLU Feed-Forward

└─ Residual Add

↓

Final RMSNorm

↓

LM Head (Linear)

↓

Logits

2.2 구성 요소 설명:

계층명	역할	구성 요소
Token Embedding	입력 문장을 벡터화	SentencePiece 기반 Subword Tokenizer
Positional Encoding (RoPE)	토큰 간 순서 정보 반영	Rotary Positional Embedding
Transformer Decoder Block	Transformer Decoder Block	RMSNorm (Pre-LN) Grouped Multi-Head Self-Attention Residual Connection RMSNorm

		SwiGLU FFN Residual Connection
Self-Attention (GQA)	장·단기 문맥 의존성 학습	Query Projection, Key/Value Projection
Feed-Forward Network (FFN)	비선형 변환을 통한 표현력 확장	Linear Projection
Language Modeling Head	다음 토큰 확률 예측	

3. 학습 과정

1. Gemma-2-2b-it와 같은 계열의 상대적 대형 모델인 gemma-2-9b-it를 LoRA를 통해 학습

항목	값
학습 데이터 수	1,014건
검증 데이터 수	200건
에폭(Epoch) 수	3
배치 크기 (Batch Size)	2
학습률 (Learning Rate)	2e-4

2. 학습된 gemma-2-9b-it 에서 토큰을 생성할 때의 logit 값을 추출
3. Gemma-2-2b-it 모델이 학습 데이터와, 이 데이터에 대한 gemma-2-9b-it 모델의 logits 값을 참고하여 teacher 모델의 target 의 확률 분포를 모방하도록 학습

4. 학습 결과 및 성능 평가

4.1. 학습 결과 요약

- BERTScore

모델 설명	Precision	Recall	F1 Score
파인튜닝을 거치지 않은 gemma-2-2b-it	0.8197	0.8709	0.8443
LoRA를 사용하여 파인튜닝한 gemma-2-2b-it	0.7905	0.7992	0.7946
LoRA를 사용하여	0.8276	0.8392	0.8332

파인튜닝한 gemma-2-9b-it			
Knowleadge Distillation을 활용하여 파인튜닝한 gemma-2-2b-it	0.8310	0.8337	0.8321

- ROUGE Evaluation

모델 설명	ROUGE1	ROUGE2	ROUGE2SUM
파인튜닝을 거치지 않은 gemma-2-2b-it	0.0894	0.0164	0.0879
LoRA를 사용하여 파인튜닝한 gemma-2-2b-it	0.3905	0.2056	0.3845
LoRA를 사용하여 파인튜닝한 gemma-2-9b-it	0.4826	0.2692	0.4797
Knowleadge Distillation을 활용하여 파인튜닝한 gemma-2-2b-it	0.4783	0.2784	0.4732

4.3 해석 및 분석

- 기존 학습 데이터와 매우 유사한 형태의 데이터 100건과, 다른 주제와 형태의 데이터 106건에 대하여 작업을 수행한 결과, Knowleadge Distillation을 활용하여 파인튜닝한 모델이 상대적으로 높은 성능을 보이는 것이 확인되었다.
- 글에서 핵심을 추출하여 요약하는 성능은 나쁘지 않았고, 같은 내용이 들어온 경우에는 거의 유사한 내용으로 텍스트를 생성하였다.

6. 결론 및 향후 계획

- 최종 선정 모델 : google/gemma-2-2b-it
- 활용 방안 : 각 텍스트 문단에 대한 임베딩용 요약 텍스트 생성
- 향후 계획:
    - 각 요약 텍스트에 대해 임베딩 후, 임베딩 결과를 확인하여 개선점 확인 및 개선
    - 모델 동작 시간 단축을 위한 경량화 예정