SK네트웍스 Family AI 과정 19기

# 데이터 전처리 학습된 인공지능 모델

| 산출물 단계 | 데이터 전처리 |
|---|---|
| 평가 산출물 | 학습된 인공지능 모델 |
| 제출 일자 | 2026-01-13 |
| 깃허브 경로 | https://github.com/orgs/Poli-Cheetah |
| 작성 팀원 | 박준영 |

1. **모델 목적**: 분할된 문서의 일부분에서 맥락에 알맞는 요약을 생성하는 모델

2. **모델 아키텍처 설계**

- 최종 선정 모델: google/gemma-2-2b-it

  모델의 선정에는 distillation을 통한 학습 가능성을 최우선으로 고려하며, 모델의 기본 성능과 사이즈를 중심적으로 고려

- 아키텍처 개요:

| 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)<br>Grouped Multi-Head Self-Attention<br>Residual Connection<br>RMSNorm<br>SwiGLU FFN<br>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 | |

- 아키텍처 시각화:

  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

3. **모델 학습 요약**

- 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 |

- 학습된 gemma-2-9b-it 에서 토큰을 생성할 때의 logit 값을 추출
- 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

- 성능 평가 결과:

  BERT Score:

| 지표 | 값 |
|---|---|
| Precision | 81.3% |
| Recall | 83.3% |
| F1 Score | 83.2% |

  ROUGE Evaluation:

| 지표 | 값 |
|---|---|
| ROUGE1 | 0.478 |
| ROUGE2 | 0.278 |
| ROUGE2SUM | 0.473 |

- 일반화 성능 평가:
  - 미검증 데이터셋(Test set)에 대한 성능 평가 결과 포함
  - 학습에 사용한 데이터, 학습에 사용한 데이터와 유사하지만 차이가 있는 데이터, 새로 생성한 데이터셋을 혼합하여 사용

4. **저장 및 배포**

- 저장 형식:

| 항목 | 설명 |
|---|---|
| 저장 파일명 | /gemma-2-final-summary-model |
| 저장 형식 | 모델의 학습 파라미터 등이 저장된 폴더 |
| 저장 방법 | student_model.save_pretrained("./gemma-2b-final-summary-model")<br>tokenizer.save_pretrained("./gemma-2b-final-summary-model") |
| 모델 불러오기 코드 예시 | model = AutoModelForCausalLM.from_pretrained(<br>student_model_path,<br>torch_dtype=torch.bfloat16,<br>device_map="auto",<br>) |

- 모델 사양 요구 사항:
    - 사용 라이브러리 : peft-0.18.1 accelerate-1.12.0 bitsandbytes-0.49.1 datasets-4.4.2 trl-0.26.2
    - GPU/CPU 호환 여부: 학습 시는 GPU 사용 필수, 사용 시에도 GPU 사용 권장
- 모델 테스트:
    - 모델 추론 테스트 완료, 적재 진행중
    - Inference 예시:

      입력: “### RAG(Retrieval-Augmented Generation)의 핵심 구성 요소\n\nRAG 시스템은 크게 세 단계로 나뉩니다. 첫째, 문서 로드 및 청킹(Chunking)을 통한 데이터 준비입니다. 둘째, 사용자 질문을 벡터화하여 관련 문서를 찾는 검색(Retrieval) 단계입니다. 셋째, 검색된 정보와 질문을 결합해 답변을 생성(Generation)하는 단계입니다.”

      출력: “RAG 시스템은 문서 로드·청킹과 함께 검색, 생성을 통해 세 단계로 구조화된다.”

5. **종합 평가 및 활용 방안**

- 일반화 가능성: 미사용 데이터셋에서도 일정 수준의 성능 유지

- 향후 활용: API 서버에 탑재하여 섹션화된 문서의 일부분을 요약하고, 기존 문서와 비교할 수 있는 데이터를 생성하는 데에 사용

6. **추가 개선 예정 사항**

- 추가 데이터셋 수집을 통한 성능 고도화
- 추론 속도 향상을 위한 모델 경량화

7. **추가 기재**

- 저장된 모델 파일 위치 또는 URL: gemma-2b-final-summary-model