SK Networks Family AI Camp 9기

**인공지능 학습 결과서**

가로선

**프로젝트 명: DAIS (Divergent AI with Science)**

**담당자: 조이현**

# <목차>

1. 모델 학습 개요

2. 워크 플로우

3. 학습 성능 분석

4. 추론 및 평가

5. 테스트 결과 및 결론

# 1. 모델 학습 개요

가. 모델 학습 목적: ​​본 모델 학습의 목적은 한국어 과학 도메인 질의응답 및 생성

태스크에서 높은 정확도를 가진 AI 인플루언서를 구축하는 것임,

또한 RunPod 환경에서 Q-DoRA 기반의 미세조정(파인튜닝)과

RAG(Retrieval-Augmented-Generation) 파이프라인을 적용하여,

과학적 질문에 대해 신뢰성 있는 답변을 생성하고, 대용량 문서 검색

및 활용 능력을 극대화하는 데 있음

나. 사용 모델 정보

1) 추론(생성) 모델: Qwen/Qwen3-8B

- 80억 파라미터, 32K 컨텍스트, instruction-aware, 최신 SOTA

(현 시점 최고 성능을 보이는 모델, 알고리듬, 기술) LLM

2) 임베딩(검색) 모델: nlpai-lab/KURE-v1

- 약 1.9B 파라미터, 8192 토큰, 1024차원,

한국어+과학/전문용어 검색 최신 SOTA

다. 모델 선정 이유

1) 추론(생성) 모델: Qwen/Qwen3-8B

가) 80억 파라미터의 대형 LLM으로, 32K 컨텍스트 지원 및 최신

instruction-aware SOTA 성능을 보유

나) 과학/기술/특허 등 다양한 이공계 도메인에 적합한 대용량 문서

처리능력과, 프롬프트 엔지니어링에 최적화된 구조

다) Q-DoRA와 같은 PEFT(파라미터 효율적 미세조정) 기법과 호환성이

우수하며, RunPod 환경에서 안정적으로 구동 가능

라) 한국어 포함 다국어 지원, Thinking Mode를 통한 추론 시스템 및

스페셜 토큰 도입과 실제 RAG 파이프라인에서 높은 생성 품질을 보임

2) 임베딩(검색) 모델: nlpai-lab/KURE-v1

가) 약 1.9B 파라미터, 8192, 1024차원으로 한국어 과학/전문용어 검색에

최적화된 최신 SOTA 임베딩 모델

나) 200만 쌍 이상의 한국어 쿼리-문서 데이터로 파인튜닝되어,

과학/기술/법률 등 다양한 전문 도메인에서 높은 검색 정확도와

Retrieval 성능을 입증

다) FAISS, Qdrant 등 VectorDB와의 연동이 용이하며, 검색 증강

생성(RAG) 시스템에서 실전 활용성이 높음

라) 벤치마크 기준 한국어 Search/Retrieval 분야에서 SOTA 성능을 기록

# 2. 워크 플로우

가. 데이터셋 준비

- 3697개 rows의 jsonl 파일, 각 row당 최대 48 tokens의 input 및

최대 2000 tokens의 output으로 구성된 과학 질의응답 데이터셋 구축

나. Qwen/Qwen3-8B 모델 파인튜닝

- Q-DoRA(Quantized Decomposition Rank Adaptation)

- 4bit 양자화, LoRA/DoRA 기반 PEFT, gradient checkpointing,

paged\_adamw\_32bit 최적화

- 주요 하이퍼 파라미터: batch\_size=32(4x8), cosine scheduler,

early stopping(patience=5)

- 프롬프트 엔지니어링: 과학 AI 인플루언서 스타일, instruction-aware,

도메인 특화 special tokens 적용, DAIS 페르소나 부여

다. 임베딩 벡터화

- nlpai-lab/KURE-v1 모델로 질문-컨텍스트 임베딩 생성,

FAISS 기반 VectorDB 구축

라. RAG 파이프라인

- FAISS 및 Retrieval로 관련 문서 검색 → Qwen/Qwen3-8B에 컨텍스트로 입력

→ 최종 답변 생성

마. 성능 평가 및 추론

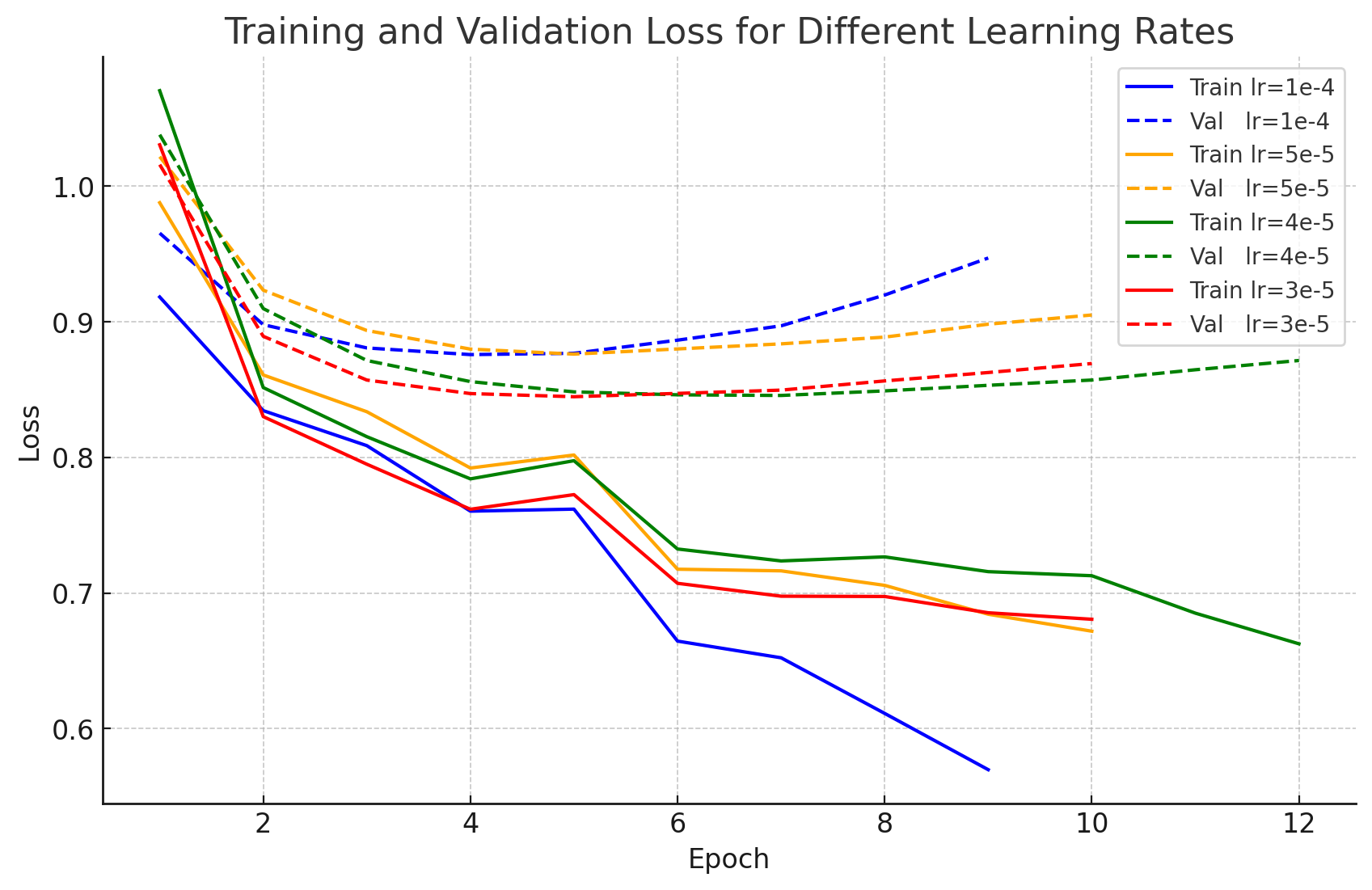
- loss 곡선, cosine similarity, perplexity, RAG 기반 답변 품질 평가

# 3. 학습 성능 분석

가. 분석 결과표

| learning\_rate | 1e-4 | 5e-5 | 4e-5 | 3e-5 |
| --- | --- | --- | --- | --- |
| Epoch별 Training Loss 변화 | 1-4 epochs  - loss 급격히 감소  4-5 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  4-9 epochs  - loss 급격히 감소 | 1-4 epochs  - loss 급격히 감소  4-5 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  5-10 epochs  - 급격히 감소 이후 완만히 감소 | 1-4 epochs  - loss 급격히 감소  4-5 epochs  - 발산 징후 보임, 이후 완만히 감소  5-12 epochs  - 급격히 감소 이후 완만히 감소 | 1-4 epochs  - loss 급격히 감소  4-5 epochs  - 발산 징후 보임, 이후 완만히 감소  5-10 epochs  - 급격히 감소 이후 완만히 감소 |
| Epoch별 Validation Loss 변화 | 1-2 epochs  - 급격히 감소  2-4 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  5 epoch  - 발산 시작, 과적합 징후 보임 | 1-2 epochs  - 급격히 감소  2-5 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  6 epoch  - 발산 시작, 과적합 징후 보임 | 1-2 epochs  - 급격히 감소  2-7 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  8 epoch  - 발산 시작, 과적합 징후 보임 | 1-2 epochs  - 급격히 감소  2-5 epochs  - 완만한 loss로 인한 안정적 수렴 구간 조금 보임  6 epoch  - 발산 시작, 과적합 징후 보임 |
| Loss간 관계에 따른 파인튜닝 적합도 | Epoch 4  - 0.875829의 최저 validation loss 관찰, 이후 5 Epochs 연속 개선 없음, early stopping 발동, 최적 모델 자동 저장 | Epoch 5  - 0.876187의 최저 validation loss 관찰, 이후 5 Epochs 연속 개선 없음, early stopping 발동, 최적 모델 자동 저장 | Epoch 7  - 0.845697의 최저 validation loss 관찰, 이후 5 Epochs 연속 개선 없음, early stopping 발동, 최적 모델 자동 저장 | Epoch 5  - 0.844744의 최저 validation loss 관찰, 이후 5 Epochs 연속 개선 없음, early stopping 발동, 최적 모델 자동 저장 |

나. 결과 사진 첨부



다. 적합도 평가: 네 가지 학습률을 시험해본 결과, 4e-5와 3e-5의 validation loss의 곡선이

가장 부드러우며, 둘 중에서 3e-5가 validation loss 최저점 보유

→ 파인튜닝 적합성이 더 높을 것으로 예상

(Source 1: “arXiv:2402.09353, Section 4, Table 6, Figure 2”, 2024, https://arxiv.org/html/2402.09353v4)

(Source 2: “AirOps, How to Fine Tune an LLM”, 2024, https://www.airops.com/blog/how-to-fine-tune-an-llm)

(Source 3: “How to fine-tune and evaluate Qwen3 with Unsloth”, 2025, https://wandb.ai/byyoung3/

Generative-AI/reports/How-to-fine-tune-and-evaluate-Qwen3-with-Unsloth---VmlldzoxMjU3OTI0Ng)

# 

# 

# 4. 추론 및 평가

가. VectorDB 구현

- FAISS는 CPU/GPU 모두 지원하며, 벡터DB 확장성과 성능이 뛰어남

- 임베딩 모델(nlpai-lab/KURE-v1)로 생성한 문서 임베딩 벡터를 FAISS

vectorstore에 저장하여 대규모 벡터 검색 인덱스를 구축

- FAISS의 IndexFlatIP(Inner Product) 인덱스를 사용하여 벡터 간 유사도를

내적(Inner Product) 방식으로 계산

- 코사인 유사도 기반 검색을 위해, 인덱스에 추가하기 전 모든 임베딩 벡터를

L2 정규화(norm=1)하여 저장

- 검색 쿼리 역시 동일하게 L2 정규화를 적용한 후, 내적을 계산하면 결과적으로

코사인 유사도와 동일한 결과를 얻을 수 있음

나. 프롬프트 엔지니어링 적용

- 과학 AI 인플루언서 스타일 및 DAIS 페르소나 부여

- instruction-aware, 과학적 개념의 쉬운 설명, 팬(사용자)과의 소통 중요시, 팬의

질의에 대해 step-back하여 의도 파악, 의도 파악 후 CoT 적용하여 논리 전개

- special tokens: [DAIS\_INSTRUCTION], [DAIS\_STYLE], [DAIS\_RULE],

[DAIS\_EXAMPLE], [HISTORY], [INPUT], [OUTPUT], [CONTEXT]

챗봇 도메인 특화 토큰 추가

다. 평가 결과표

| 지표 | 벡터스토어 A | 벡터스토어 B | 우위 |
| --- | --- | --- | --- |
| Cosine Similarity (avg) | 0.6880 | 0.6880 | 동일 |
| Perplexity (avg) | 2.2197 | 2.2086 | B (낮을수록 좋음) |
| BLEU-4 (avg) | 0.5342 | 0.5374 | B (높을수록 좋음) |
| ROUGE-L (avg) | 0.3829 | 0.3824 | A (높을수록 좋음) |
| 응답 시간 (avg, sec) | 139.0559 | 129.7084 | B (짧을수록 좋음) |

라. 적합도 평가

- 코사인 유사도는 A·B 모두 0.6880으로 동일해, Top-1 문서 검색 성능 차이 없음

- Perplexity는 B(2.2086)이 A(2.2197)보다 약간 낮아, B의 예측이 더 안정적

- BLEU-4는 B(0.5374) > A(0.5342)이나, 차이 극미

- ROUGE-L은 A(0.3829) > B(0.3824)이나, 차이 극미

- 응답 시간은 B가 평균 129.7초로 A(139.1초) 대비 약 9초 빠름

- 종합 판단: 실사용 측면(속도)과 전반적 문장 일치도(BLEU-4 및 낮은

Perplexity)를 고려하면 벡터스토어 B가 더 잘 형성되었다고 볼 수 있음,

ROUGE-L은 A가 아주 약간 앞서지만 전체적으로는 B가 우세함

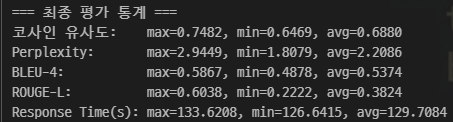
# 5. 테스트 결과 및 결론

가. 실제 질의응답 테스트

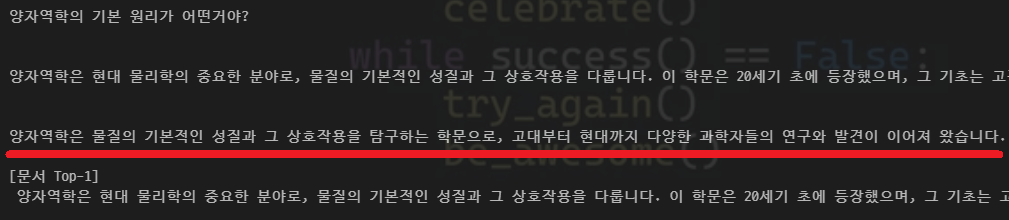
- 과학/특허/기술 질문에 대해 맥락 있는 답변 생성

- RAG 파이프라인 적용 시, 외부 문서 근거 기반 답변 품질 향상

나. 정량적 지표



다. 정성적 평가



라. 결론

- Qwen3-8B + Q-DoRA + KURE-v1 + FAISS 기반 RAG 파이프라인으로

한국어 과학 도메인에서 SOTA 수준의 생성·검색·RAG 품질을 달성했음

- 정량·정성 평가 모두에서 실전 적용에 충분한 성능과 안정성을 입증했음