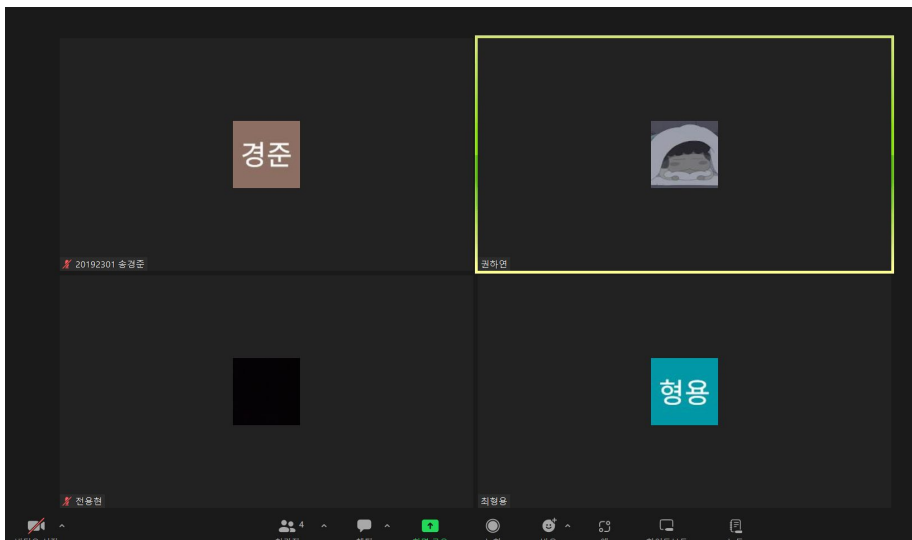


CUAI 프로젝트 NLP 1팀

2024.07.02

발표자 : 전용현

스터디원 소개 및 만남 인증



스터디원 1 : 권하연
 스터디원 2 : 송경준
 스터디원 3 : 전용현
 스터디원 4 : 최형용

목차

Review

Fine Tuning

RAG

Streamlit을 이용한 챗봇 구현

Review

“QA 모델링을 통한 CUI 챗봇 만들기”

선정 배경

: CUI의 데이터 확인의 어려움

=> 이를 취합하여 정리 후 QA 챗봇을 생성

프로젝트 목표

- RAG의 활용으로 LLM의 한계를 극복한다
- QA 모델링을 활용해본다

구현 예시

Q : 2023년 하반기에 NLP 3팀이 한 프로젝트를 알려줘

A : Diffusion 기반 Text Generation

Review

Data

처리 전/후 데이터 모두 활용

INDEX

Retrieval

Augment

- 청크와 **overlap** 사이즈 조정
- **chunker** 변경

- 리트리버 모델 탐색
- 하이퍼파라미터 조정

- **generation**

병합진행

+ Fine tuning

Fine Tuning

Model

MLP-KTLim/llama-3-Korean-Blossom-8B

미세조정 기법

LoRA (Low-Rank Adaptation)

lora_alpha: 16

- LoRA의 스케일링 파라미터. 학습의 안정성을 조절.

r: 128

- 낮은 랭크 행렬의 차원.
- 높은 차원은 더 많은 파라미터를 의미하며, 이는 더 높은 성능을 가짐.
- 낮은 차원은 파라미터 수를 비약적으로 줄일 수 있어 한정된 리소스 내 작동이 가능케 함

target_modules: ['q_proj', 'down_proj', 'up_proj', 'o_proj', 'v_proj', 'k_proj', 'gate_proj']

- 미세 조정할 특정 모듈을 지정.

RAG

Chunker

RecursiveCharacterTextSplitter

embedder

OpenAIEmbeddings

Retriever

BM25Retriever
FAISS
EnsembleRetriever

Generator

gpt-3.5-turbo

Implement

streamlit

RAG - Chunker

RecursiveCharacterTextSplitter

- 재귀적 문자 텍스트 분할 : **의미적으로 관련 있는** 텍스트 조각들이 같이 있도록 설계
- `chunk_size`와 `chunk_overlap`를 조정해가며 **최적의 답변**을 내놓는 값 도출
- **dataset의 형태**에 따른 모델의 성능 평가
 - QA vs text
 - text 형식일 때 모델의 성능이 더 좋아짐을 확인

```
# chunk documents
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 500,
    chunk_overlap = 50,
    length_function = len)
texts = text_splitter.split_documents(data)
```

- `chunk_size` : 각 청크의 **크기** 제한(500)
- `chunk_overlap` : 청크 간 문자의 **중첩** 허용 정도(50)
- `length_function` : `len` 함수를 활용 -> 텍스트의 **길이**를 계산

RAG - Retriever

ensemble_retriever = BM25 + FAISS

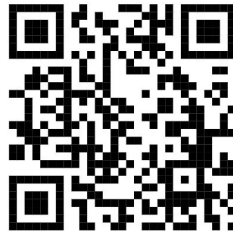
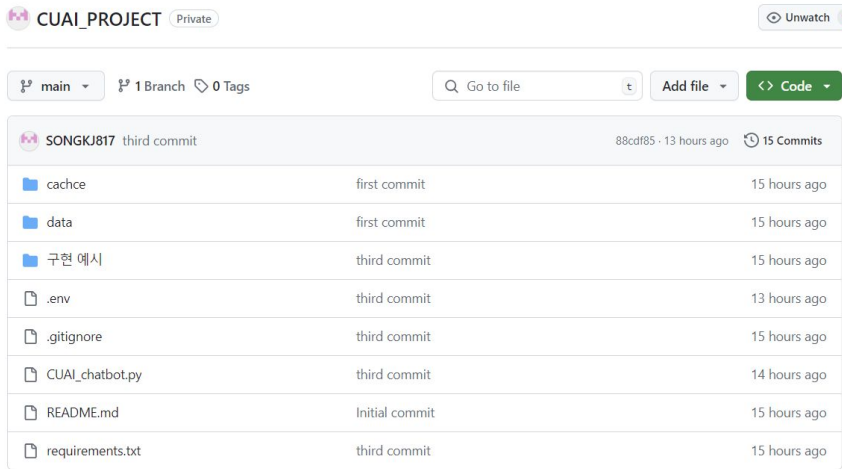
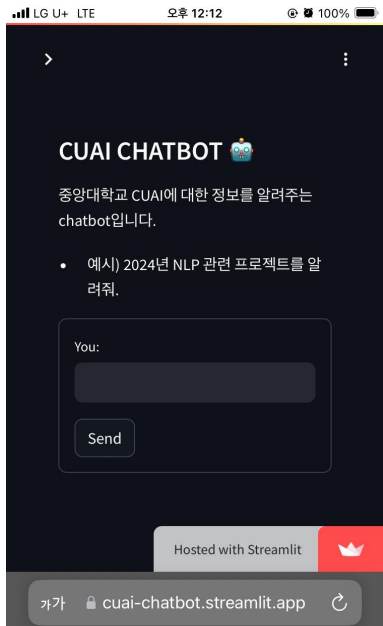
- 각각의 retriever에서 2개의 document를 가져오도록 함
- BM25: Sparse Retriever - TF-IDF를 고려하여 질의와 문서 간의 단어 매칭 빈도 검색
- FAISS: Dense Retriever - 사전 정의된 embedding을 이용하여 벡터 공간 내 근접 이웃 검색
- BM25는 키워드 매칭, FAISS는 의미적 유사성 매칭에 강점을 가지고 있음
- 둘의 장점을 동시에 활용함으로써 전체적 맥락과 키워드가 비슷한 document를 검색할 수 있음

```
def generate_response(prompt):  
    llm = ChatOpenAI(model_name = "gpt-3.5-turbo", temperature=0)  
    handler = StdOutCallbackHandler()  
    qa_with_sources_chain = RetrievalQA.from_chain_type(  
        llm=llm,  
        retriever=ensemble_retriever,  
        callbacks=[handler],  
        return_source_documents=True  
    )  
    response = qa_with_sources_chain({"query" : f"{prompt}"})  
    return response["result"]
```

RAG - Streamlit

Streamlit

- streamlit을 통해 구현 + github과 연동하여 URL 배포



CHATBOT 구현 예시

CUAI CHATBOT

중앙대학교 CUI에 대한 정보를 알려주는 chatbot입니다.

- 예시) 2024년 NLP 관련 프로젝트를 알려줘.

You:

Send

2023년 CV 관련 프로젝트를 알려줘



2023년 CV 관련 프로젝트로는 CUI CV T2(반도체) 프로젝트가 있습니다. 이 프로젝트 팀은 SNN(Spiking Neural Network) 이론을 학습하고 논문을 리뷰하여 저전력 SNN 시스템을 사용하여 컴퓨터 비전 분야에 적용하는 작업을 진행했습니다. 이 프로젝트는 SNN 모델을 사용하여 MNIST 데이터셋에 대한 초기 테스트를 진행하고 차중 식별 이미지 처리 모델을 구현했습니다. SNN은 기존 FNN 모델에 비해 높은 성능을 보여주었으며, 학습률 스케줄러 사용으로 성능 향상이 이루어졌습니다.

2021년 NLP 관련 활동이 있다면 알려줘



2021년 NLP 관련 활동으로는 CUI 4기 NLP 2팀이 국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여 다양한 머신러닝 및 딥러닝 모델로 MBTI 성격 유형을 예측하는 연구를 진행했습니다. 해당 연구에는 Future of NLP + Deep Learning 등의 주제가 포함되어 있었습니다.

CUAI에 대해 알고 있어?



네, CUI는 Chung-Ang 대학교 인공지능 연구회의 약자로, 인공지능 및 머신러닝 분야에 관심 있는 학생들이 모여 다양한 스터디 그룹을 통해 지식을 공유하고 함께 성장하는 활동을 하는 조직입니다.

안녕?



안녕하세요! 무엇을 도와드릴까요?

2024년 CV 관련 프로젝트를 알려줘



죄송합니다. 2024년 CV 관련 프로젝트에 대한 정보는 제가 알고 있는 범위를 벗어납니다.

의의 & 한계점

의의: 1. CUI의 활동 자료를 살펴보고 싶은 사람들을 위한 최초의 챗봇

2. 최신 연구인 **RAG**를 사용하여 트렌디한 챗봇 구현

3. **Streamlit**을 사용하여 사용자에게 친숙한 인터페이스 구축

4. CUI 자료가 갱신됨에 따라 챗봇의 자동 업데이트

한계점 : 1. 챗봇을 전문적인 평가 지표를 가지고 직접적으로 평가하지 않음. (몇 번의 **inference**를 통해 눈으로 체크)

2. 데이터셋의 규모가 너무 작아 **fine-tuning** 효과 x. 추후에 규모가 크고 더욱 정제된 데이터셋으로 가공 시 **fine-tuning**까지 결합할 수 있음을 기대