DAT 4기 캡스톤 프로젝트 보고서

LLM 을 활용한 사용자 맞춤 음식 추천 시스템 구현

: 이미지 기반 음식명 추출 및 대화형 추천을 중심으로

DAT 4기

A4

김동영

김동휘

김형수

황서진

Data Analysis & Technology

DAT 4기 캡스톤 프로젝트 보고서

LLM 을 활용한 사용자 맞춤 음식 추천 시스템 구현

: 이미지 기반 음식명 추출 및 대화형 추천을 중심으로

Development of a Personalized Food Recommendation System Using LLM

: Focusing on Image-Based Food Name Extraction and Conversational Recommendations

이 보고서를 캡스톤 프로젝트 보고서로 제출합니다.

2024년 11월 30일

DAT 4기

A4

김동영

김동휘

김형수

황서진

Data Analysis & Technology

DAT 4기 캡스톤 프로젝트 보고서

LLM 을 활용한 사용자 맞춤 음식 추천 시스템 구현

: 이미지 기반 음식명 추출 및 대화형 추천을 중심으로

국문 요약

본 프로젝트는 사용자가 업로드한 음식 사진과 사용자 선호 정보를 기반으로 개인화된 음식 추천 시스템을 설계하는 데 중점을 두었다. 이미지에서 음식명을 추출하기 위해 BLIP 모델을 활용하였으며, GPT 모델을 통해 사용자 선호도에 적합한 추천 음식을 생성하였다. 이를 통해 사용자는 자신이 선호하는 요리를 쉽고 효율적으로 추천 받을 수 있으며, 레시피까지 제공받을 수 있는 시스템을 구현하였다. 해당 시스템은 음식 선택 고민을 해소하고, 창의적인 요리 아이디어를 제공하는 데 실질적인 기여를 할 것으로 기대된다.

Abstract

This study focuses on designing a personalized food recommendation system based on images of food uploaded by users and their preference information. The BLIP model was utilized to extract food names from the images, and the GPT model was employed to generate recommended dishes that align with the user's preferences. Through this system, users can easily and efficiently receive food recommendations tailored to their tastes, along with recipe suggestions. The system is expected to provide practical contributions by alleviating the dilemma of food selection and offering creative cooking ideas.

DAT 4기

A4

김동영

김동휘

김형수

황서진

Data Analysis & Technology

목 차

그림 목차

제 1 장 서론

제 1 절 연구 배경

최근 음식 문화는 단순히 식사를 넘어 개인의 선호와 감정을 반영하고, 이를 타인과 공유하는 경험으로 확장되고 있다. 특히 SNS 의 발달은 이러한 경향은 더욱 가속화 시켰다. 사람들은 음식을 만드는 과정이나 소비한 음식을 사진 또는 영상으로 기록하며 이를 공유하고, 이러한 활동은 하나의 문화 콘텐츠로 발전했다. 그러나 음식의 종류와 선택지가 지나치게 다양해지면서 오히려 무엇을 먹을지 결정하지 못하는 경우가 빈번하게 발생하고 있다. 이에 따라 "무엇을 먹을까?"라는 질문은 단순한 선택의 문제가 아니라, 시간과 에너지를 소비하는 고민으로 이어지고 있다.

최근 대규모 언어 모델(LLM)의 발전은 이미지 및 텍스트 데이터를 분석하고 유의미한 정보를 제공할 수 있는 새로운 가능성을 열고 있다. 이러한 기술을 바탕으로 사용자 중심의 맞춤형 음식 추천 시스템을 설계하고자 한다. 특히 사용자가 촬영한 음식 사진이나 최근 섭취한 음식 목록, 개인의 선호도를 기반으로 추천을 생성하는 알고리즘은 현대인의 바쁜 일상 속에서 효율적인 음식 선택을 지원하는 데 유용할 것이다. 이러한 시스템은 사용자가 보다 간편하게 음식 선택 고민을 해결하고, 나아가 새로운 음식과 요리에 대한 창의적인 아이디어를 제공할 것으로 기대된다.

제 2 절 연구 목적 및 논문 구성

본 프로젝트는 OpenAI GPT 모델과 Python 을 활용하여 사용자 데이터를 기반으로 대화형 음식 추천 및 요리 레시피 제공 시스템을 설계하고 구현하는 것을 목표한다. 절차는 다음과 같다.

1. 사용자가 제공한 음식 사진을 분석하여 음식 이름을 추출하고, 이를 CSV 파일에 저장한다.

2. 추출된 음식 목록을 바탕으로 사용자의 선호정보를 조사한다. 선호정보는 재료, 조리 시간, 매운맛 여부 등 여러 요인으로 구성된다.

3. 수집된 데이터를 기반으로 LLM 모델을 활용하여 사용자의 취향에 맞는 음식 추천과 요리 레시피를 작성한다.

이러한 시스템은 사용자가 일상에서 느끼는 "무엇을 먹을까?"라는 고민을 덜어주고, 동시에 새로운 요리 아이디어를 제공함으로써 요리에 대한 흥미와 창의성을 자극할 수 있다. 이미지와 텍스트 데이터를 분석하여 유의미한 정보를 도출하는 LLM 은 음식 사진에서 이름을 추출하고, 사용자 선호를 반영한 맞춤형 추천을 제공함으로써 사용자 중심의 혁신적인 서비스를 구현할 수 있다. 이를 통해 음식 선택 과정에서의 불편함을 해소하고, 사용자의 최근 경험과 선호도를 기반으로 효율적이고 개인화된 추천을 제공하는 이 시스템은 현대인의 바쁜 일상에 실질적인 가치를 더할 것으로 기대된다.

제 2 장 모델

본 프로젝트에서는 음식 이미지에서 음식 이름 텍스트를 추출하고 사용자가 입력한 정보를 받아 이 두 정보를 융합하여 음식 메뉴와 레시피를 추천해주는 모델을 구상하였다.

먼저 구체적인 process를 설명하기에 앞서 프로젝트에서 사용한 AI architecture에 대해 소개하고자 한다.

Image에서 Text를 추출하기 위해 BLIP 모델을 사용하였다.

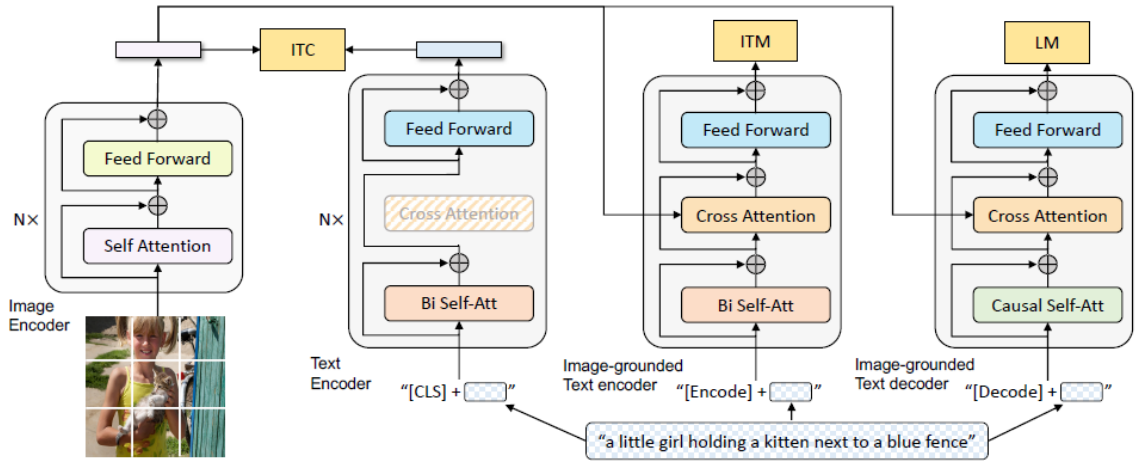


그림 2-1. BLIP 모델 구조

출처: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

제 1 절 BLIP

BLIP은 Bootstrapping Language-Image Pre-training의 약자로 2022년 Salesforce에서 개발한 Multi Modal model이다. Image-Text retrieval, Image captioning, Visual question answering, Visual reasoning 등 다양한 Vision-Language task에서 높은 성능을 자랑한다.

BLIP architecture는 Unimodal encoder, Image-grounded text encoder, Image-grounded text decoder로 나눠져있다. 각 파트의 특징을 소개하자면 먼저 Unimodal encoder는 Image와 Text를 독립적으로 encoding을 하는데 Text는 BERT, Image는 ViT를 encoder로 사용한다. Text encoder는 Image에 대한 Text feature를 담고 있는 [CLS] Token을 Text input 앞에 추가한다. Image-grounded Text encoder는 Transformer block에서 Self-attention layer와 Feed-

forward network 사이에 Cross Attention layer를 삽입했다. Cross Attention layer의 역할은 Image feature와 Text feature 간의 deep interaction을 파악하기에 BLIP에서 핵심을 맡고 있다. Image-grounded Text decoder는 Bidirectional Self-Attention layer를 Casual Self-Attention layer로 대체하였다. Causal Self-Attention은 현재 Token의 생성에 이전 Token만 참고하도록 제한하여 Text sequence 구조를 유지하면서 생성해내는 역할을 맡고 있다.

Image-grounded Text decoder는 시퀀스의 시작을 알리는데 사용되는 [Decoder] Token을 Text input 앞에 추가한다.

이름에서 알 수 있듯이 BLIP은 대규모 Image-Text 데이터셋으로 사전 학습하여 일반화 성능을 올리는 방법을 사용한다. 다음으로 BLIP Pre-training 방식인 Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), Language Modeling Loss (LM)에 대해 간단히 소개하고자 한다.

Image-Text Contrastive Loss (ITC)는 unimodal encoder를 활성화하여 Positive Image-Text pair가 유사한 representation을 갖도록 학습하는 방식이다.

Image-Text Matching Loss (ITM)은 Image-grounded text encoder를 활성화해 Image-Text pair가 일치한 지를 예측하는 이중 분류 작업이다.

Language Modeling Loss (LM)은 Image-grounded Text decoder를 활성화해 Image를 조건으로 올바른 text를 생성하는 것을 목표로 학습이 진행된다.

Text 에서 Text 를 생성하기 위해 GPT 모델을 사용했다.

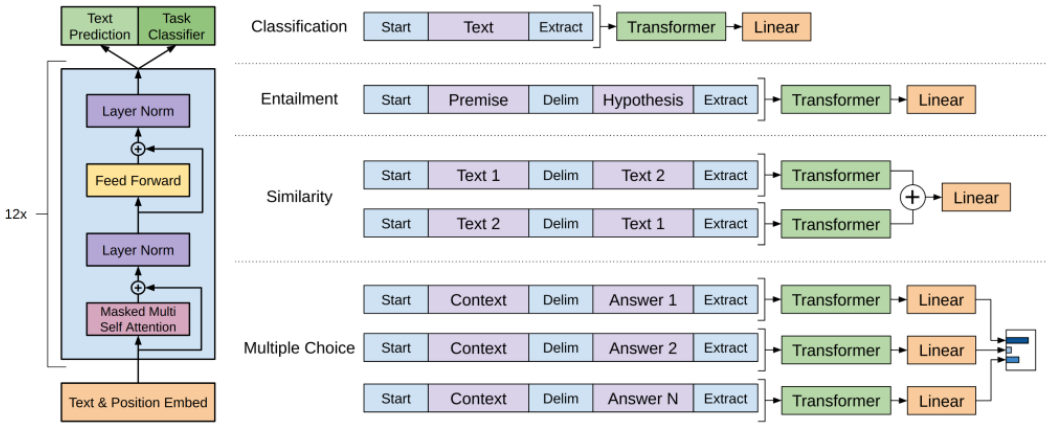


그림 2-2. GPT 모델 구조

출처: Improving Language Understanding by Generative Pre-Training

제 2 절 GPT

GPT 모델은 Transformer Decoder architecture를 기반으로 한다. 주요 특징으로는 Self-Attention을 통해 Token 간의 문맥적 관계를 파악하고 Feedforward Layer를 통해 비선형 변환을 추가하여 표현력을 강화한다. 또한 Causal Masking을 사용하여 학습 중 모델이 미래 Token을 참조하지 못하도록 제한하며, 왼쪽에서 오른쪽으로의 생성 sequence을 준수 한다.

GPT 모델의 Unsupervised pre-training은 다음 단어를 맞추도록 진행된다. GPT의 Unsupervisied pre-training은 대규모 Unlabeled corpus를 활용하여 다양한 downstream task가 용이하다. Pre-Training을 마친 모델은 최종 성능 측정을 하기에 앞서 각 Task에 맞도록 재학습을 한다. 이 과정을 Fine Tuning이라고 하는데 Fine Tuning은 앞선 Pre-Training과 달리 Supervised Learning 방법으로 이루어진다.

제 3 장 요리 추천 시스템 구현

제 1 절 데이터 전처리

3.1.1 자료수집

추천을 위한 분석에 사용된 자료는 사용자가 직접 업로드한 음식 사진 데이터를 이용하였다. 사용자는 자신이 최근에 맛있게 먹은 음식 사진을 5 장 업로드하고, 이를 통해 추천 시스템에서 음식을 분석하도록 하였다. 음식 사진은 LLM API 를 사용해 텍스트 형태의 음식명으로 변환된다. 데이터는 이후 사용자의 음식 선호도를 바탕으로 추천 시스템에서 활용된다. 해당 최근 맛있게 먹은 음식 정보와 이후 추천 받은 요리 중 사용자가 선택한 요리정보와 레시피 정보는 firebase 에 저장되어 추후 요리 추천에 반영 가능하다.

3.1.2 이미지 데이터 전처리

이미지 데이터는 먼저 512x512 해상도로 리사이즈된 후, Base64 형식으로 변환되어 GPT-4o 모델을 통해 분석되었다. 웹 환경에서 사용자가 업로드한 이미지는 XFile 객체로 처리되며, 이를 다양한 방식으로 접근할 수 있게 해준다. 이 XFile 객체는 다시 Base64 형식으로 변환되어 GPT-4 모델에 전송되고, 모델은 이미지를 분석하여 음식의 이름을 추출한다. 이렇게 변환된 음식 이름은 사용자가 최근에 즐긴 음식으로 기록되며, 추천 시스템에서 사용자 맞춤형 추천을 위해 활용된다. 추출된 음식 정보는 사용자의 최근 취향을 반영한 것으로 간주되며, 이를 바탕으로 음식 추천이 이루어진다.

제 2 절 LLM 을 통한 음식 추천

3.2.1 대화형 LLM 을 통한 음식 추천

LLM 을 활용한 음식 추천 시스템은 사용자가 업로드한 이미지에서 추출된 음식명과 사용자가 직접 선택한 음식 선호 정보를 기반으로 음식을 추천해주는 방식으로 구현되었다. 이 시스템에서는 GPT-3.5-turbo 모델을 사용하여 사용자의 현재 선호정보와 이전에 즐겼던 음식을 바탕으로 음식 추천을 생성한다.

음식 추천 프로세스는 다음과 같다:

1. 사용자가 응답한 선호도 정보를 바탕으로 사용자 취향 분석.
2. 사용자가 업로드한 이미지에서 음식명 추출.
3. GPT-3.5-turbo 모델에 최근 먹은 음식 목록과 선호 정보를 전달하여 5 개의 추천 요리를 생성.

사용자는 “주 식재로”, “알러지가 있는 재료”, "매운 정도", "온도", "요리 시간", "식사 종류(아침, 점심 등)"과 같은 정보를 선택하여 시스템에 전달하며, 이러한 선택 정보는 음식 추천에 반영된다.

3.2.2 LLM 을 통한 음식 추천 방식

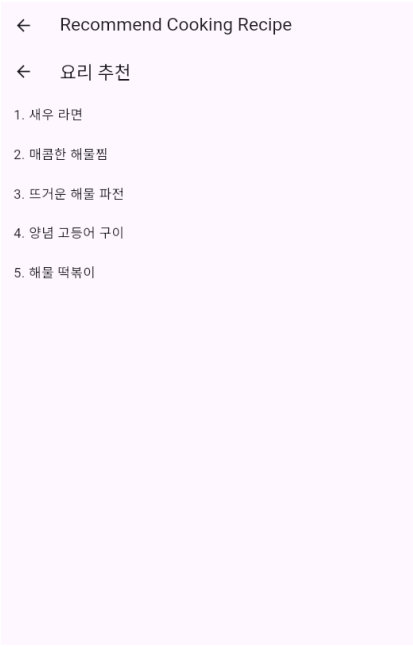
GPT-3.5-turbo 모델은 사용자가 업로드한 이미지에서 추출한 음식명과 사용자가 선택한 선호도 정보를 바탕으로 최적의 음식을 추천한다. 추천 요청 시 "최근에 먹은 음식"과 "선호도 정보"를 프롬프트로 제공하여 GPT-3.5-turbo 가 사용자의 상황에 맞는 5 가지 추천 요리를 제안한다.

이 과정에서, 시스템은 대화형 모델의 특성을 활용하여 단순한 정보 제공을 넘어 사용자가 더욱 선호할 만한 음식을 제안하고, 그에 대한 레시피까지도 제공할 수 있다. 사용자는 제안된 음식 중 하나를 선택하여 레시피를 제공받을 수 있으며, 이는 챗봇과의 자연스러운 대화를 통해 이루어진다.

제 3 절 구현 결과 및 특징

3.3.1 음식 추천 결과 예시

(a)



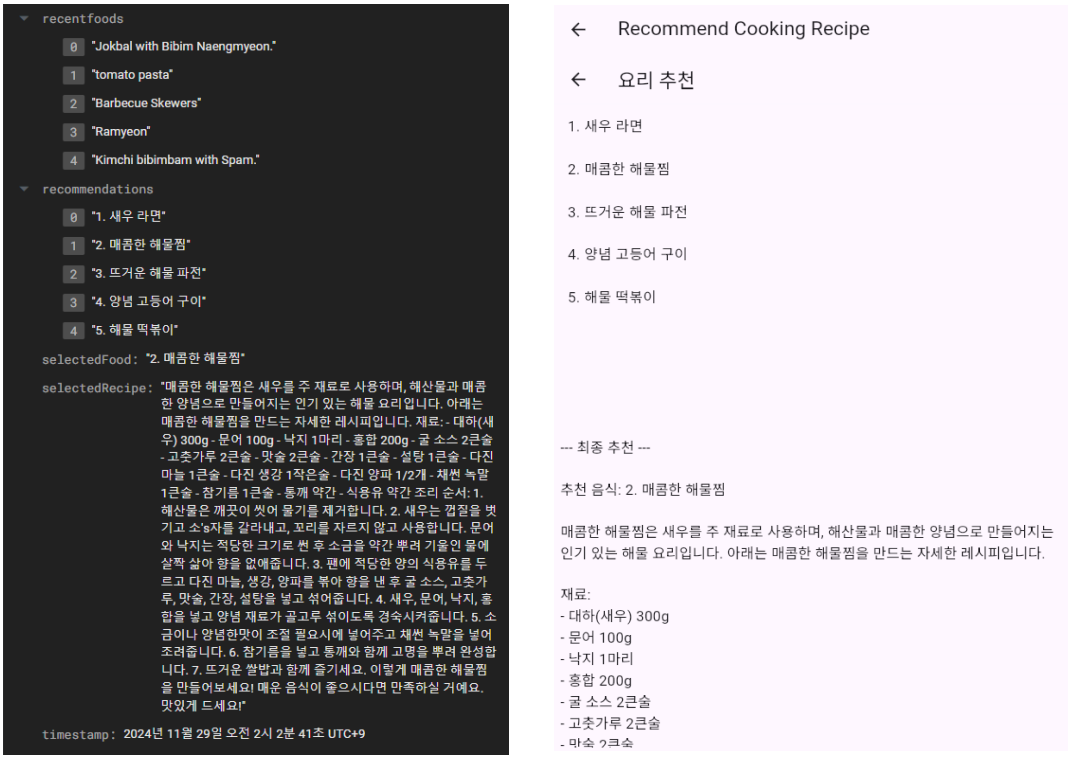
(b)

그림 3-1. (a) 사용자 현재 선호도 및 최근 맛있게 먹은 음식 사진 (b) 추천받은 요리 목록

사용자가 업로드한 음식 이미지를 기반으로 음식 이름을 추출한 후, 이를 바탕으로 GPT-3.5-turbo 를 사용하여 5 개의 추천 요리를 생성했다. 사용자가 그림(1)과 같이 “새우”, ”땅콩”, "매운거”, “뜨거운거”, "30 분 이내", "저녁 식사"이라는 선호도를 설정한 경우, 추천 시스템은 다음과 같은 음식들을 제안한다.

1. 새우 라면
2. 매콤한 해물찜
3. 뜨거운 해물 파전
4. 양념 고등어 구이
5. 해물 떡볶이

사용자가 이중에서 매콤한 해물찜을 선택하면, 해당 음식의 재료와 요리 방법을 그림(2)와 같이 안내해준다.



(a) (b)

그림 3-2. (a) Firebase 에 저장된 정보 (b) 사용자가 선택한 요리 재료 및 레시피

사진 이미지를 텍스트로 변환하는 작업은 firebase 에 저장된 정보를 토대로 각 이미지와 맞는 텍스트로 변환된 것을 확인할 수 있고 매콤한 해물찜의 경우 처음에 사용자가 주 식재료로 선택한 새우를 재료로 사용하는 모습을 볼 수 있다. 이를 통해 사용자의 현재 선호 및 최근 맛있게 먹었던 음식정보를 요리 추천 시스템이 잘 반영하는 것으로 볼 수 있다.

제 4 장 결론

제 1 절 연구 결과

본 프로젝트는 사용자가 업로드한 음식 사진 데이터에서 LLM API 를 활용해 음식명을 추출하고, 사용자가 입력한 선호 정보를 결합하여 GPT-3.5-turbo 를 활용해 개인화된 음식 추천

및 레시피 제공 서비스를 구현했다. 또한, 이전에 추천 받은 음식 기록을 조회할 수 있는 기능을 통해 서비스의 지속적인 활용성과 보다 개인화된 서비스를 제공한다.

제 2 절 연구 한계점 및 제언

본 프로젝트는 사용자가 업로드한 음식 사진에서 추출한 음식과 선호 정보를 바탕으로 맞춤형 음식 추천 및 레시피 제공 서비스를 구현했다. 본 프로젝트의 한계점은 다음과 같다. 첫째, 사용자가 제공하는 선호 정보가 한정적이다. 현재 시스템은 서비스에서 제공하는 질문에 한정하여 사용자가 선택한 선호 정보를 바탕으로 음식을 추천하는 방식이기 때문에, 복잡한 취향을 충분히 반영하지 못할 가능성이 존재한다. 둘째, 사용자가 추천 받은 음식에 대한 평가나 피드백을 제공할 수 있는 기능이 부족하여 사용자의 선호를 학습하고 이를 기반으로 시스템을 개선하는데 제약이 있다. 마지막으로, 음식에 대한 텍스트 설명이나 영양 정보와 같은 추가적인 데이터를 활용하지 않는다.

이러한 한계점을 보완하기 위해 몇 가지 개선점을 제안한다. 첫째, 사용자가 제공하는 선호 정보를 더욱 세밀하게 반영하기 위해, 추가적인 질문을 통해 사용자의 기분이나 상태를 파악하고 이를 추천 과정에 반영할 수 있다. 또한, 계절별 음식이나 지역별 특산물 등을 포함한 다양한 추천 옵션을 추가하여 사용자의 취향을 더욱 폭넓게 반영할 수 있는 방안을 모색해야 한다. 둘째, 사용자가 추천 받은 음식을 평가하고 피드백을 제공할 수 있는 시스템을 도입하여, 이 피드백을 바탕으로 추천 시스템을 지속적으로 개선할 수 있다. 마지막으로, 음식에 대한 설명 텍스트나 영양 정보 등 추가적인 정보를 분석하여 추천의 정확성을 높이는 방법도 고려할 수 있다.

참고문헌

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
Improving Language Understanding by Generative Pre-Training