캡스톤디자인 I 계획서

(표지)

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td colspan="4">Generation VL 모델을 활용한 Multi-Modal ChatBot 어플리케이션</td></tr>
<tr><td>영문</td><td colspan="4">Multi-Modal ChatBot Application using Generation VL Model</td></tr>
<tr><td>프로젝트 목표
(500자 내외)</td><td colspan="5">우리는 최첨단 VL(Vision Language) 생성 모델을 사용하여 텍스트와 이미지를 모두 처리할 수 있는 멀티모달 챗봇 애플리케이션을 만드는 것을 목표로 하고. 챗봇은 사용자의 입력을 기반으로 응답을 생성하고, 필요에 따라 관련 정보나 이미지에 대한 답변을 제공할 수 있게 된다.</td></tr>
<tr><td>프로젝트 내용</td><td colspan="5">우리는 기존의 챗봇 어플리케이션과의 차이점으로 더 나은 성능을 위해 ‘시각적 추론’이라는 개념을 제안한다. 시각적 추론은 시각적 정보를 사용하여 문제를 해결하거나 질문에 답하는 능력을 말한다. 여기에는 이미지나 비디오와 같은 시각적 데이터에 존재하는 관계, 속성 및 구조에 대한 이해와 추론도 포함된다. 우리는 시각적 추론이라는 개념이 기존의 VL모델에 유의미한 성능향상을 도출해 낼 것이라 가정하고 실험하여 증명 할 것이다. 일련의 과정들은 학술대회나 저널에 제출하여 엄중하고 객관적인 심사를 받게 할 것이고. 성능향상이 증명된 VL 모델로 챗봇 어플리케이션을 웹 서비스를 통해 제작하여 실질적으로 일반적인 사용자가 사용할수 있겠금 한다.</td></tr>
<tr><td>중심어(국문)</td><td>비전-언어 모델</td><td colspan="2">시각적 추론</td><td>챗봇</td><td>웹 서비스</td></tr>
<tr><td>Keywords (english)</td><td>Vision-Language model</td><td colspan="2">Visual Reasoning</td><td>Chatbot</td><td>Web service</td></tr>
<tr><td>멘토</td><td>소속</td><td colspan="2">서울과학기술대학교</td><td>이름</td><td>임경태</td></tr>
<tr><td rowspan="4">팀 구성원</td><td>학년/반</td><td>·학 번</td><td>이 름</td><td colspan="2">연락처(전화번호/이메일)</td></tr>
<tr><td>4</td><td>20181602</td><td>최창수</td><td colspan="2">choics2623@gmail,com</td></tr>
<tr><td>4</td><td>20181632</td><td>임현석</td><td colspan="2">gustjrantk@gmail.com</td></tr>
<tr><td>4</td><td>20201752</td><td>이현서</td><td colspan="2">20201752@edu.hanbat.ac.kr</td></tr>
</table>

컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.

2023 년 03 월 02 일

책 임 자 : 최창수 (인)

희망 지도교수 : 장한얼

(내용)

1. 캡스톤디자인의 배경 및 필요성

1) 수행하려는 프로젝트 과제와 관련되는 국내·외 연구, 산업 현황, 문제점 및 전망 등에 관하여 기술

1. 국내 연구: 한국의 IT 기업들이 자연어 처리 기술과 인공지능 기술을 활용한 챗봇 서비스를 개발하고 있습니다. 예를 들면, 네이버의 Clova AI와 카카오 AI 등이 있습니다. 또한 대학 연구기관들도 챗봇에 관한 연구를 진행하고 있습니다.
2. 해외 연구: 해외의 대표적인 기업들로는 구글, 아마존, 마이크로소프트 등이 있으며, 이들 기업은 인공지능 기술과 자연어 처리 기술을 활용하여 챗봇 서비스를 제공하고 있습니다. 또한 대학 연구기관들도 챗봇에 관한 연구를 진행하고 있습니다.
3. 산업 현황: 챗봇은 고객상담, 예약, 주문, 검색 등의 분야에서 활용되고 있으며, 많은 기업들이 챗봇을 도입하여 업무효율을 높이고 있습니다. 특히 최근에는 멀티모달 챗봇이 등장하면서, 이미지나 비디오 등의 멀티미디어 자료에 대한 처리 능력이 강화되고 있습니다.

기존의 chatbot의 문제점

1. 제한된 대화 능력: 대부분의 기존 chatbot은 단순한 rule-based나 keyword-based 방식으로 작동하여, 다양한 대화 상황에 대응하지 못하고 반복적이고 지루한 대화를 제공합니다.
2. 인식 능력 부족: 일부 chatbot은 사용자의 발화를 완전히 이해하지 못하거나 의도를 파악하지 못하는 경우가 있습니다. 이는 대화의 효율성을 떨어뜨리고 사용자 경험을 나쁘게 합니다.
3. 정보 제공 한계: 대부분의 chatbot은 단순한 텍스트 응답을 제공하며, 이미지나 비디오 등의 멀티미디어 자료에 대한 대응 능력이 부족합니다.
4. 지속적인 개선 필요: chatbot은 학습 데이터의 양과 질에 크게 영향을 받으며, 새로운 데이터나 대화 상황에 대한 대응 능력을 강화하기 위해 지속적인 개선이 필요합니다.

2) 프로젝트의 필요성을 구체적으로 기술

우리의 제안법(VL+VR)은 기존의 챗봇의 문제점 중 3개(제한된 대화 능력, 인식 능력 부족, 정보 제공 한계)대한 극복 방안으로 제안할수 있다.

1. 제한된 대화 능력(rule-based, keyword-based)은 GPT 계열의 Generation 모델을 활용하여 다양성을 확보, 반복적이지 않으며 사용자가 원하는 대답을 할수 있게한다.
2. 인식 능력 부족은 Visual reasoning의 관계적이고 유추적인 정보를 VL 모델에 추가하여 사용자의 발화에 대한 이해를 도와 의도를 파악할수 있게 도움을 줌.
3. 정보 제공의 한계 극복은 기존의 uni-modal로 이루어진 챗봇(chatGPT, 카카오플러스 등)과 다르게 Vision Language 모델을 활용하여 text to text, image to text로 표현이 가능해진 multi-modal로 극복을 할 수 있다.

위 세가지 이유로 캡스톤 디자인으로 본 프로젝트의 필요성을 이야기한다.

2. 캡스톤디자인 목표 및 비전

시각 질의 응답을 위해 '그림에 대해서 설명해줘'라는 질문에 답하기 위해서는 질문, 이미지에 대한 내용이 필요하다. Visual reasoning의 정보를 VL 모델에 추가한 모델을 만들고자 한다. VR은 Contrastive learning을 사용할 것이다. 데이터셋은 대규모 시각 추론 데이터 240,000장(gettyimages), Pre-train VL 데이터를 사용할 예정이다.

3. 캡스톤디자인 내용

주요 기능

기능	내용
데이터 분석	- matplotlib, seaborn으로 데이터 시각화 - 지식기반을 전처리하기 위하여 pandas 라이브러리 활용
질의 모델	- pytorch 기반의 gpt-2 모델 사용 - 한국어, 영어가 있으므로 각 언어마다 모델에 맞게 Fine-tuning
이미지 모델	- pytorch 기반의 VIT + CNN계열 모델 사용
VR	- contrastive learning을 활용하여 정보 추출

주요 기능

	내용
성능	- 기존 gpt-2 모델의 성능과 VL + VR을 활용한 모델 성능 비교 - VR 정보 활용 방안을 연구하여 성능 향상
보안	- 논문을 통해 외부에 모델 공개
결과	- GPT-2 모델을 사용하여 사용자에게 시각 질의 응답 서비스 제공 - 연구한 모델을 바탕으로 논문 작성

4. 캡스톤디자인 추진전략 및 방법

대규모 시각 추론 데이터는 240,000 개의 데이터로 이루어져 있으며 대용량 데이터 셋이다. 데이터 셋 학습은 VIT 계열 모델과 Contrastive loss를 사용하기 위해서 GPU와 64GB이상의 서버 메모리가 필요하기에 서버를 활용하여 실험한다. 서버는 서버 메모리 128GB, GPU NVIDIA A100 80GB X 4이다.

2022년 대규모 시각 추론 Visual Reasoning 과제를 착수, 데이터 셋에 대한 깊은 이해도를 가지고 있다. 또한 'GPT-2' 모델 및 'VIT', 'Contrastive loss'에 대한 이해가 있으며 연구실 활동을 통해 공부한 자연어처리 모델을 이해하고 사용할 수 있다.

서울과학기술대학교에서 MLP 연구실을 이끌고 있는 임경태 교수를 멘토로 섭외하였다. 임경태 교수는 자연어처리 기반의 멀티모달 러닝의 전문가이다.

	팀 구성	성명	주요 역할
1	팀장	최창수	논문 및 자료조사, 모델 코드 작성
2	팀원	임현석	논문 및 자료조사, 모델 코드 작성
3	팀원	이현서	자료조사 및 UI 개발

사용 프레임워크

5. 참고문헌

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[2] Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

[3] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

[4] Johnson, Justin, et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.