캡스톤디자인 II 계획서

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td colspan="4">Visual Reasoning Information을 활용한 한국어 기반 LLM 웹 어플리케이션</td></tr>
<tr><td>영문</td><td colspan="4">Korean based LLM Web Application using Visual Reasoning Information</td></tr>
<tr><td>프로젝트 목표
(500자
내외)</td><td colspan="5">ChatGPT가 세계적으로 큰 관심을 끌며 LLM(Large Language Model : 대규모 언어 모델)의 중요도가 높아지고 있다. 하지만 기존의 LLM은 한국어 기반의 LLM이 많지 않다. 그렇기에 기존의 LLM 프롬프트에 한국어를 입력하게 되면 원하지 않는 답변이 나오는 경우가 많다. 기존의 LLM은 단어와 문장을 일련의 통계적 패턴으로 이해하고 생성하지만 한국어는 주어와 목적어의 생략, 조사의 다양한 사용 등 문맥에 대한 특성에 있어 이를 올바르게 이해하고 처리하는 것이 어렵다. 또한 한국어는 다른 언어에 비해 많은 어휘 다양성을 가지고 있다. 많은 단어들이 유사한 뜻을 가지고 있고, 동음이의어나 다의어의 사용이 빈번하다. 이에 우리는 한국어를 기반으로 한 LLM을 제안한다. 또한 캡스톤 디자인1에서 입증한 VR(Visual Reasoning) Information을 활용하여 사전훈련된 VR Model과 LLM을 연계하는 align 모델을 만들 것이다.</td></tr>
<tr><td>프로젝트
내용</td><td colspan="5">우리의 궁극적인 목표는 한국어 기반 LLM을 만들고 해당 LLM에 Pre-training VR Model을 연계하여 웹 어플리케이션으로 배포하는 것이다. 우리는 ‘KULLM’, ‘Koalpaca’ 데이터셋을 통해 학습시킨 한국어 기반 LLM은 다양한 어휘 및 문맥의 특성을 이해하고 프롬프트의 입력 대비 사용자가 원하는 출력을 얻을 것이다. 그리고 모델의 대규모 파라미터 업데이트에 대항하여 LoRA(Low-Rank Adaption of Large Language Models) fine-tuning 기법을 사용한다. 이 후, 해당 LLM은 웹 어플리케이션으로 배포하기 때문에 다양한 사용자가 직접 사용이 가능하다.</td></tr>
<tr><td>기대효과
(500자
이내)
(응용분야 및
활용범위)</td><td colspan="5">우리의 모델은 한국어에 대해 자연스러운 답변과 생성을 가능케할 것이고 이를 통하여 유용하고 맞춤화된 정보와 답변을 얻을 수 있다. 또한 한국어로 된 다양한 주제에 대한 질문, 설명, 가이드 등을 제공하여 학습과 정보 습득에 도움을 줄 수 있을 것이다.</td></tr>
<tr><td>중심어(국문)</td><td>대규모 언어 모델</td><td>시각적 추론</td><td>구름</td><td colspan="2">대규모 언어
모델의 하위 적용</td></tr>
<tr><td>Keywords
(english)</td><td>LLM</td><td>Visual Reasoning</td><td>KULLM</td><td colspan="2">LoRA(Low-Rank
Adaption of Large
Language Models)</td></tr>
<tr><td>멘토</td><td>소속</td><td colspan="2">서울과학기술대학교</td><td>이름</td><td>임경태</td></tr>
<tr><td rowspan="5">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">01057182623 / choics2623@gmail.com</td></tr>
<tr><td>4</td><td>20181632</td><td>임현석</td><td colspan="2">01089300790 / gustjrantk@gmail.com</td></tr>
<tr><td>4</td><td>20201752</td><td>이현서</td><td colspan="2">01051096689 / 20201752@edu.hanbat.ac.kr</td></tr>
<tr><td colspan="6">컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.
2023 년 6 월 22 일
책 임 자 : 최창수 (인)
지도교수 : 장한얼 (인)</td></tr>
</table>

캡스톤디자인 계획서(양식)

1. 캡스톤디자인의 배경 및 필요성

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

2) 프로젝트의 필요성을 구체적으로 기술

ChatGPT가 세계적으로 큰 관심을 끌며 LLM(Large Language Model : 대규모 언어 모델)의 중요도가 높아지고 있다. 하지만 기존의 LLM은 한국어 기반의 LLM이 많지 않다. 그렇기에 기존의 LLM 프롬프트에 한국어를 입력하게 되면 원하지 않는 답변이 나오는 경우가 많다. 기존의 언어모델인 ChatGPT, Bing과 같은 경우 다국적 언어모델이다. 서구권 언어에 비해 Low-Resource 언어인 한국어와 같은 언어들은 서구권 언어를 기반으로 cross-lingual transferability로서 한국어를 처리하기 때문에 비교적 한국어 성능이 떨어지게 된다. 따라서 한국어를 위한 LLM모델의 필요성이 대두되었다. 'Polyglot-ko', 'KoAlpaca', 'KoVicuna', 'KULLM'와 같이 한국어 특화 모델이 제안되었고, 본 캡스톤 프로젝트의 목표로 가장 성능이 특화된 한국어 특화 LLM 모델을 제작할 것이다.

2. 캡스톤디자인 목표 및 비전

기존의 LLM은 단어와 문장을 일련의 통계적 패턴으로 이해하고 생성하지만 한국어는 주어와 목적어의 생략, 조사의 다양한 사용 등 문맥에 대한 특성에 있어 이를 올바르게 이해하고 처리하는 것이 어렵다. 또한 한국어는 다른 언어에 비해 많은 어휘 다양성을 가지고 있다. 많은 단어들이 유사한 뜻을 가지고 있고, 동음이의어나 다의어의 사용이 빈번하다. 이에 우리는 한국어를 기반으로 한 LLM을 제안한다. 또한 캡스톤 디자인1에서 입증한 VR(Visual Reasoning) Information을 활용하여 사전훈련된 VR Model과 LLM을 연계하는 align 모델을 만들 것이다.

3. 캡스톤디자인 내용

1) 캡스톤디자인의 주요 기능, 비 기능적 요구사항(성능, 보안, 유지보수성 등)

본 캡스톤 디자인에서는 'KULLM', 'Koalpaca'의 데이터 셋을 활용하여 한국어 기반 LLM을 생성한다. 모델은 polyglot으로 연구를 진행 후 우리가 생성한 LLM을 바탕으로 진행한다. 이 후, 대용량 파라미터 업데이트에 대항하여 LoRA로 fine-tuning을 진행한다. 이후 캡스톤 디자인1을 바탕으로 Pre-trained된 Visual Reasoning 모델과 우리가 만든 한국어 기반의 LLM을 연계하는 align 모델을 최종적으로 생성한다.

해당 모델을 바탕으로 우리는 웹 어플리케이션을 통해 배포한다. 해당 웹 어플리케이션에서는 회원가입 및 로그인 기능을 제공하며 프롬프트에 사용자가 한국어 입력을 제공하면 모델을 통하여 사용자가 원하는 출력을 도출한다.

우리가 사용하는 프레임워크는 Pandas, Pytorch, Hugging Face를 통해 모델을 구축한다. Flask를 사용하여 웹 어플리케이션 구축, Jinja2 템플릿 엔진을 통하여 동적 컨텐츠를 생성하고 jQuery를 사용하여 서버 응답에 답하며 MongoDB를 사용하여 데이터베이스를 관리한다.

2) 추진일정 등을 기술

월	6		7						8					9					10					11				
주	4	5	1	2	3	4	5	6	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
모델 연구	o	o	o	o																								
모델 생성					o	o	o	o	o	o	o	o	o	o	o	o												
실험 진행																	o	o	o	o	o	o	o					
웹 서버 구축														o	o	o	o	o	o	o								
웹 모델 연결																					o	o	o					
최종 검토																								o	o	o	o	o

4. 캡스톤디자인 추진전략 및 방법

1) 캡스톤디자인 목표 달성을 위한 추진전략, 수행방법 및 추진절차를 기술함

*** 참고: 추진전략의 작성 ****

(1) 캡스톤디자인에 대한 이해
캡스톤디자인 추진을 위한 예상 문제점을 식별하고 프로젝트 추진을 위한 준비방안 수립

(2) 캡스톤디자인 경험
저학년 프로젝트를 통해 습득한 기본 기술 활용 및 Lessons Learned를 활용

(3) 검증된 멘토 활용
관련된 경험과 역량을 풍부하게 축적하고 있는 멘토를 활용

(4) 프로젝트 관리체계 수립
역할 및 책임을 명확히 하도록 프로젝트 관리 체계를 수립

자연어 데이터는 동음이의어, 문장의 다양성, 언어의 다양성과 같은 이유로 문제점이 많다. 해당 문제점을 해결하기 위해 데이터 전처리에 많은 신경을 써야 한다. 우리는 오픈 데이터셋인 'KULLM' 데이터셋을 사용한다. GPU와 64GB이상의 서버 메모리가 필요하기에 서버를 활용하여 실험한다. 서버는 서버 메모리 128GB, GPU NVIDIA A100 80GB X 4이다.

현재 TeddySum과 협업하여 LLM 과제를 착수하고 있어 많은 연구 및 사전조사를 진행중에 있다. 또한 'KULLM'. 'Koalpaca' 데이터 셋에 대해 분석 중에 있으며 연구실 활동을 통해 공부한 자연어처리 실력을 기반으로 모델을 이해하고 사용할 수 있다.

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

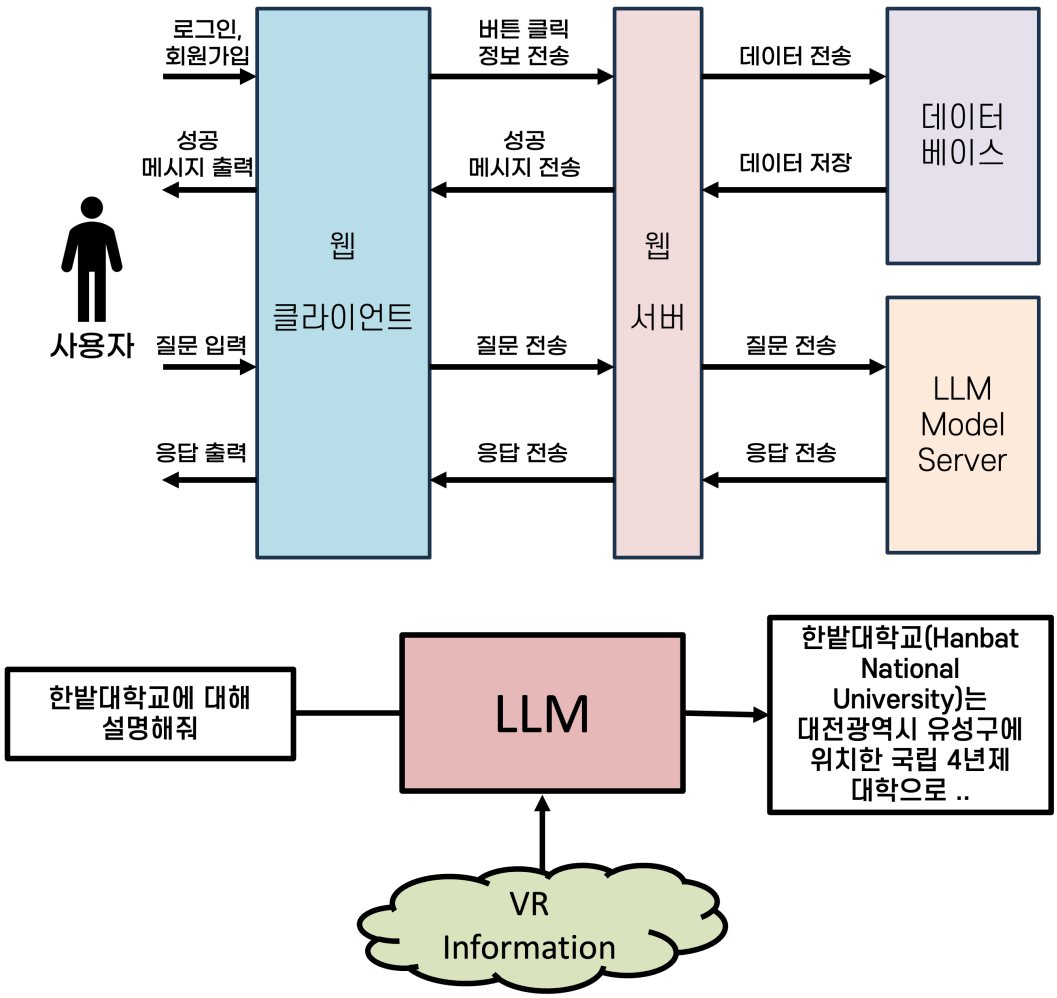
*** 참고: 수행방법의 작성 ****

캡스톤디자인 **수행을 위한 방법론, 프레임워크, 분석툴에 대해 작성**

2) 캡스톤디자인 목표 달성을 위한 팀 구성 체계 및 역할에 대하여 기술함

	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



<수행방법>

위의 그림과 같은 프롬프트의 입력을 받고 출력이 가능한 모델 개발을 우선 순위에 둔다. 먼저 'KULLM'. 'Koalpaca' 데이터셋을 기반으로 하여 LoRA fine-tuning을 적용시킨 모델을 생성하며 보다 가벼운 모델을 개발한다. 개발한 모델을 기반으로 로컬에서 사용 가능한 웹 어플리케이션을 서버로 구축하여 모델과 연결하여 최종적으로 한국어 기반 ChatBot 어플리케이션을 개발할 것이다.

5. 캡스톤디자인 결과의 활용방안

1) 제안된 캡스톤디자인이 추진되었을 경우의 사회적/기술적/경제적 파급효과 등을 자유롭게 기술함

LLM은 'ChatGPT'의 등장으로 큰 화두에 올랐지만 한국어 기반의 LLM은 많지 않다. 실제로 'ChatGPT'의 프롬프트에 한국어로 작성할 시, cross-lingual transferability로서 한국어를 처리하기 때문에 비교적 한국어 성능이 떨어지게 되는데 한국어 기반의 LLM은 해당 문제점을 해소할 수 있을 것이다. 한국어 기반의 LLM을 개발한다면 한국어의 특성상 동음이의어, 다양한 문장 형태를 잘 파악하여 텍스트를 요약하거나 문서를 생성해주는 등 기존의 LLM 역할을 한국어에 맞춰 높은 성능을 보일 것으로 예상된다.

6. 참고문헌

[1] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language

models." arXiv preprint arXiv:2106.09685 (2021).
[2] Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).

캡스톤디자인 II 계획발표 채점표

	학년/반	학 번	이 름
팀 구성원	4	20181602	최창수
	4	20181632	임현석
	4	20201752	이현서
제 목			

항목	점수				
	1	2	3	4	5
1. 프로젝트 주제의 필요성이나 중요성이 적절히 서술되었는가?					
2. 국내외 동향(문제 제기), 주요 기능(특징 포함) 및 범위가 적절히 서술되었는가?					
3. 기대효과(사회적, 기술적, 경제적 파급효과)가 적절히 서술되었는가?					
4. 추진 전략과 수행방법이 적절한가?					
5. 팀 구성과 역할 분담이 적절히 이루어졌는가?					
합계					

*수정 및 개선 의견

2013년 월 일

심사위원 : (인)

※ 채점은 각 영역별 5점 만점을 기준으로 채점함.(상 5, 중 3, 하 1)

※ 계획서와 발표내용을 참고하여 채점표에 따라 평가함.