|  |  |
| --- | --- |
| **프로젝트 개발보고서**  -프로젝트 명- | |
|  | |
| 성 명 | 이한울 |
| 프로젝트명 | Llama with Iphone |

1. 모델 구축 개요

□ 배경 및 필요성

○ 프로젝트 착수 배경

대형 언어 모델(LLM)이 일상에 깊이 스며들면서 개인정보 유출과 같은 윤리적 문제가 부각되고 있다. 서버 기반 모델은 사용자 데이터가 외부로 전송되고, 학습 데이터로 활용될 가능성을 내포하며, 이는 민감한 정보 보호와 사용자 신뢰에 심각한 영향을 미친다.

특히, 프라이버시에 민감한 사용자들에게 안전하고 신뢰할 수 있는 AI 솔루션의 필요성이 대두되었다. 이에 따라 데이터 처리와 모델 추론을 장치 내에서만 수행하는 **On-Device AI** 기술을 활용하여, 개인정보 보호와 AI 활용을 동시에 만족시키는 방안을 모색하게 되었다.

○ 문제정의

최근 OpenAI의 ChatGPT와 같은 대형 언어 모델(LLM)이 일상생활의 일부로 자리 잡으면서 개인정보 유출 문제가 점차 부각되고 있다. 이러한 문제는 사용자의 대화 기록이 OpenAI 서버로 전송되고, 잠재적으로 학습 데이터로 활용될 가능성에서 기인한다. 이는 민감한 데이터의 외부 유출과 프라이버시 침해 우려를 야기하며, 사용자 신뢰를 약화시키는 요인으로 작용하고 있다.

○ 문제 해결을 위해 기술 스택

1. CoreML
   1. Apple이 제공하는 머신러닝 프레임워크로, Meta의 Llama 모델을 변환 및 최적화하여 iOS 기기에서 실행.
2. Pytorch & Llama
   1. Llama는 Meta에서 공개한 언어 모델(LM)로, 변환 후 On-Device 환경에 적합하도록 최적화
   2. CoreML 형식으로의 변환을 위한 도구로 활용.
3. Swift & SwiftUI
   1. iOS 환경에서의 사용자 인터페이스(UI) 구현.

2. 프로젝트 진행 상세 내용

□ 모델 처리

○ 모델 전처리 수행 과정

모델을 Meta에서 공개한 아키텍처 그대로 변환시 연산이 느려지거나 변환시 오류가 발생한다. 이를 위해서 모델의 아키텍처를 변환에 적합하게 일부 수정한다.

○ 모델 변환

모델을 CoreML 프레임워크를 통해서 변환 및 양자화를 진행한다.

□ 앱 개발

○ 사용할 SDK 설명

Apple의 iPhone 등의 Native 앱 개발 SDK인 Xcode 와 SwiftUI,MLTensor를

사용하게 된다.

○ 앱개발 단계

Figma를 통한 앱 디자인을 통해 앱을 구상하고 SwiftUI로 앱 디자인을 구성 및 LM을 구동하는 객체를 생성하여 앱을 완성하였다.

3. 프로젝트 수행 결과

□ 앱 개발 결과

○ AI 모델 성능

디바이스의 성능별로 모델의 성능이 다르게 나왔으나 프레임워크별 속도가 다르게 나왔다. 이러한 이유로는 Apple 의 IOS의 영향이 큰것으로 보이며 이를 추가적인 조사를 통해 개선의 필요성을 느낌.

○ 구축한 AI 모델을 통해 해결가능한 문제

On-Device의 구동은 성공적이였으며 vectorDB등 여러 기술을 결합하여

다양한 상황에서 구동될 수 있는것을 확인하였다.

○ AI 모델 제한점 및 문제점, 추후 해결 방안

여전히 추론 성능이 문제이다. Llama3.2 3b 모델로 실험한 결과

Iphone 16 pro 기준 2 token/s 속도로 너무 낮은 속도가 나온다.

다양한 방면에서 모델의 추론속도를 높일 수 있도록 연구를 해보고 잘 안될시

Metal 등을 통해서 프레임워크를 구축할것이다.

□ 개발일정

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 업무  (예시) | 6월 | | | | | | |
| 10일 | 11일 | 12일 | 13일 | 14일 | 17일 | 18일 |
| 주제 선정 |  |  |  |  |  |  |  |
| 문제 정의 |  |  |  |  |  |  |  |
| 데이터 수집 |  |  |  |  |  |  |  |
| 개발 기획서 작성 |  |  |  |  |  |  |  |
| 데이터 전처리 |  |  |  |  |  |  |  |
| EDA |  |  |  |  |  |  |  |
| 모델 구축 |  |  |  |  |  |  |  |
| 모델 평가 및 모델 성능 개선 |  |  |  |  |  |  |  |
| 결과보고서 작성 |  |  |  |  |  |  |  |
| 발표자료 작성 |  |  |  |  |  |  |  |
| 프로젝트 보고서 작성 |  |  |  |  |  |  |  |