DeepwaveSLM 앱 소개

DeepwaveSLM은 사용자의 글자 입력을 받아 OnDevice 환경에서 SLM 기술을 활용하여 실시간으로 글자를 출력하는 안드로이드 앱입니다. 이 앱의 목적은 비행기 모드에서도 사용할 수 있으며, 목적에 따라 RAG로 벡터임베딩된 SLM을 사용할 수 있게 하는 것입니다.

앱의 주요 도전과제

모델의 크기와 성능

언어 모델은 많은 자원을 소비합니다. 이를 모바일 디바이스에서 실행 가능하도록 최적화하는 것이 큰 도전 과제였습니다

데이터 타입 불일치

TensorFlow Lite 모델을 사용할 때 자주 발생하는 문제 중 하나는 데이터 타입 불일치입니다. 특히, INT32와 FLOAT32 사이의 변환 문제가 발생했습니다.

빌드 및 종속성 문제

Gradle 설정 및 종속성 관리 문제로 인해 빌드 과정에서 여러 차례 오류가 발생했습니다.

시연 작동 사진 및 영상

Google Docs

Screen Recording 2024-06-13 at 8.47.55 PM.mov

개발 과정의 시행착오

1

호출 및 복잡성 오류 회피

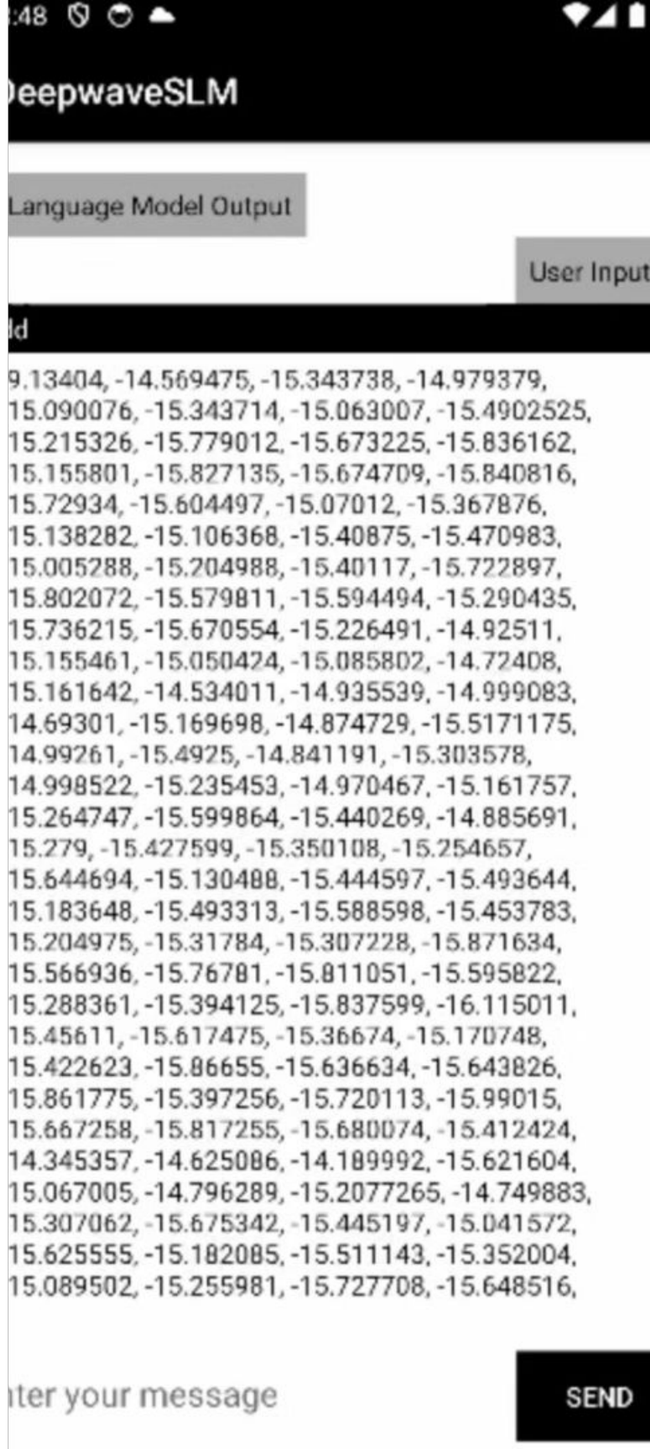
처음에는 이미 구현되어 있는 예제 LLM android 앱을 적극적으로 참조하려고 하였습니다

그러나 알 수 없는 종속성 및 라이브러리 호출 오류, 지나친 복잡성 등의 문제로 인해 앱을 처음부터 다시 설계는 방향으로 진행했습니다.

2

모델의 크기와 성능 최적화

대형 언어 모델은 많은 자원을 소비합니다. 이를 모바일 디바이스에서 실행 가능하도록 최적화하는 것이 큰 도전 과제였습니다.



개발 과정의 시행착오

1

데이터 타입 불일치 문제

INT32와 FLOAT32 사이의 변환 문제가 발생했습니다. 처음에는 INT32 타입의 입력 데이터를 사용하여 모델을 실행하려고 했습니다. 그러나 TensorFlow Lite 모델은 FLOAT32 타입을 요구했기 때문에 타입 불일치 오류가 발생했습니다. FLOAT32 포맷으로 재교육하고, 코드에서 데이터 타입 변환을 명확하게 처리했습니다. 그러나 알 수 없는 데이터 변환 오류가 지속되었습니다

2

Gradle 설정 및 빌드 문제

처음에는 기본 Gradle 설정을 사용하여 프로젝트를 빌드하려고 했습니다. 그러나 필요한 종속성이 제대로 설정되지 않아 빌드 오류가 발생했습니다. 프로젝트 수준과 앱 수준의 build.gradle.kts 파일을 수정하여 필요한 종속성을 명확하게 정의했습니다. 반복된 디버깅 작업에서 많은 시간과 리소스를 소모했습니다.

개발 과정의 시행착오

1

Send 버튼 작동 문제

처음에는 Send 버튼이 정상적으로 작동했으나, 이후에 작동하지 않는 문제가 발생했습니다. 버튼 클릭 이벤트를 확인하고 디버깅을 통해 문제를 해결하려고 했습니다. 그러나 원인을 파악하기 어려웠습니다. 코드와 로그를 면밀히 분석한 결과, 데이터 타입 변환과 관련된 문제가 버튼 작동에 영향을 미친 것을 확인했습니다.

2

모델 호환성 문제

BERT 모델을 TensorFlow Lite 형식으로 변환하는 과정에서 호환성 문제가 발생했습니다. TensorFlow Lite Converter를 사용하여 모델을 변환하는 과정에서 옵션을 조정하고, 모델 최적화 기술을 적용하여 호환성 문제를 해결했지만 오류가 지속되었습니다. Phi2B_cpu, Falcon1B_cpu등의 모델을 .bin형태로 사용하려고 하였으나 지속되는 오류로 인해 실패하였습니다.

벡터 임베딩 및 **RAG**시연

```
Question: What can I do to prepare for a certification exam after being away for two weeks?
Context: If you've been away from your certification exam preparation for two weeks, it's important to get back on track quickly. Start by reviewing your notes and st
Answer: [CLS] what can i do to prepare for a certification exam after being away for two weeks ? [SEP] if you ' ve been away from your certification exam preparation
```

Google Docs

52AM.mov

개선사항

1 모델 경량화

새롭게 출시되는 모델의 성능과 용량을 지속적으로 모니터링하고, 필요시 추가적인 최적화 작업을 수행합니다.

2 데이터 타입 불일치 문제

설계 이전 단계부터 데이터타입 변환 과정을 미리 파악하고 목적에 맞는 전처리 과정과 SLM제작이 필요합니다.

3 모델 테스트 강화

개발 기간과 개발에 필요한 자원을 적절히 세팅해야합니다. 언제 개발이 완료될지 알 수 없으므로 지속적인 모델 테스트 및 지속적인 정보 업데이트를 받아야 합니다.

4 연구 필요성

해당 프로젝트는 구현 자체로 의미가 있음을 알게 되었습니다. 현재 해당 기술을 구현할 수 있는 개발자 및 정보가 매우 희소합니다.