**토마토 스마트팜 특화 소형 언어모델의 효율적 경량화 연구**

**Efficient Lightweighting of Tomato Smart Farm-Specific Small Language Model**

1.서론

 전 세계적으로 농업에서는 생산성 향상과 노동력 절감을 위해 스마트팜 기술이 확산되고 있으며, 특히 토마토 재배는 온도, 습도, CO₂ 농도, 광량 등 환경 조건에 따라 수확량과 품질이 크게 달라 데이터 기반 관리가 중요하다. 그러나 현장에서는 전문 지식 부족과 복잡한 데이터 해석 문제로 어려움을 겪는 경우가 많고, 대규모 언어모델(LLM)을 활용한 지원은 연산 비용과 메모리 요구 때문에 농가에서 실용성이 낮다. 본 연구에서는 Gemma 1B 모델을 기반으로 가중치 프루닝, LoRA 파인튜닝, 양자화를 적용하여 라즈베리파이 CPU 환경에서도 10초 이내 응답 가능한 경량화 챗봇 시스템을 제안하며, 제한된 자원 환경에서 실시간 대화형 AI 적용 가능성을 보여주고 스마트팜 현장에서 활용 가능한 저비용·고효율 AI 보조 시스템의 기반을 마련하고자 한다.

2.관련연구

스마트팜(Smartfarm)

스마트팜은 농업 분야에 정보통신기술(ICT)를 접목해 작물의 생육환경을 원격 및 자동으로 제어할 수 있는 시설농업기술을 의미한다. 특히 농업생산 과정에서 데이터(Data)와 인공지능(AI)를 비롯한 첨단기술을 기반으로 의사결정함으로써 농업의 생산성과 효율성을  극대화한다.

현재 상용화된 스마트팜 시스템은 온도, 습도, CO2 농도, 광량과 같은 생육 환경 데이터를 센서로 수집하여 모니터링하고, 빅데이터와 인공지능, 사물인터넷을 활용해 정밀한 생육관리를 한다.

그러나 현재의 스마트팜 시스템에는 여전히 몇 가지 한계점을 지닌다.

첫째, 단순 규칙 기반 접근의 한계이다.  기존 스마트팜 제어 시스템은 주로 사전에 정의된 임계값과 규칙에 따라 환경을 제어한다. 예컨대 온도, 습도, CO₂ 농도, 광량이 특정 범위를 벗어나면 알람을 제공하거나 장치를 가동하는 방식이다. 그러나 이러한 단순 규칙 기반 접근은 복합적인 원인 해석에는 취약하다. 실제 농업 환경에서는 여러 변수들이 동시에 상호작용하여 생육 문제를 발생시키며, 특정 센서값만으로는 문제의 본질을 설명하기 어렵다. 이에 본 연구에서 제안하는 “규칙 기반-소형 언어모델(sLM) 하이브리드 기법”은 환경 이상 발생 시 그에 따른 영향과 조치 방법에 대한 응답을 해주면서도, 필요한 경우 복합 원인 해석으로 확장될 수 있다. 예를 들어 센서를 통해 측정되는 스마트팜 내부 온도가 높음과 동시에 습도가 낮은 경우, 토양 건조 위험에 따라 관수의 필요성을 응답한다.

둘째, 인터넷과 계산자원의 한계이다. 많은 스마트팜 시스템은 안정적인 네트워크 연결과 클라우드 연산 자원에 크게 의존한다. 그러나 농촌 지역에서는 인터넷 환경이 불안정하거나, 클라우드 서비스 이용에 따른 비용과 지연(latency)이 문제가 되기도 한다. 이로 인해 실시간 제어 및 의사결정 지원에 제약이 발생한다. 본 연구가 제안하는 경량 모델 기반 로컬 추론(on-device inference)은 이러한 한계를 완화할 수 있는 방안을 제공한다. 즉, 대규모 모델 전체를 사용하는 대신, LoRA 기반 경량 파인튜닝과 양자화를 통해 소규모 장치에서도 빠르고 효율적인 추론을 가능하게 하여, 인터넷 연결이 원활하지 않은 환경에서도 안정적인 서비스 제공이 가능하다.

이러한 한계점을 극복하기 위해 최근 인공지능 기술, 특히 대규모 언어 모델(Large Language Models, LLMs)과 프루닝을 이용한 경량화, 특정 도메인 학습을 위한 LoRA 파인튜닝, 양자화 기술들은 본 연구에서 제안하는 접근 방식의 기반이 되고 있다.

3. 방법론

본 연구에서는 토마토 스마트팜 환경에 적합한 소형 언어모델(sLM)을 구현하기 위해 \*\*데이터 기반 구조적 프루닝(Data-driven Structured Pruning)\*\*과 \*\*LoRA(Low-Rank Adaptation)\*\*를 결합한 경량화 전략을 채택하였다.

**구조적 프루닝(Structured Pruning)**: 모델의 가중치를 개별 노드나 커넥션 단위가 아닌 필터, 채널, 헤드 등의 구조 단위로 제거하여 연산량을 줄이고 모델 희소성을 높인다. 특히 본 연구에서는 토마토 스마트팜 센서 로그와 질의응답 데이터를 기반으로 중요도가 낮은 구조 단위를 선택적으로 제거하여 도메인 지식을 유지하면서 불필요한 연산을 줄였다.

**LoRA 기반 파인튜닝(Low-Rank Adaptation Fine-Tuning)**: 기존 언어모델의 주요 파라미터는 고정(freeze)하고, 일부 층에 저차원 행렬을 추가하여 도메인 특화 학습을 수행한다. 이를 통해 메모리 사용량과 연산 비용을 최소화하면서도 도메인 적응 성능을 유지할 수 있다.

**양자화(Quantization)**: 최종적으로 경량화된 모델 전체에 4bit 양자화를 적용하여 메모리 사용량을 줄이고, 라즈베리파이 CPU 환경에서도 실시간 추론이 가능하도록 최적화하였다.

결과적으로, 본 접근 방식은 **도메인 지식 유지 + 연산량 감소 + 저자원 환경 구동 가능성**을 동시에 달성하는 것을 목표로 한다.

4.제안 방법

본 연구는 토마토 스마트팜에서 수집되는 센서 데이터를 활용하는 **온디바이스 소형 언어모델(sLM) 경량화**를 목표로 한다. 실험에는 Gemma3-1B 모델을 사용하였으며, 라즈베리파이 CPU 환경에서도 실시간 응답이 가능하도록 다음과 같은 경량화 파이프라인을 적용하였다.

1. **데이터 기반 구조적 프루닝**:

약 5,000개의 센서 로그 및 질의응답 데이터를 활용하여 중요도가 낮은 구조 단위를 제거함으로써 모델 연산량을 감소시킴.

1. **LoRA 기반 파인튜닝**:

원본 모델 파라미터는 고정하고 일부 저차원 행렬만 학습하여 도메인 특화 성능을 유지하면서 메모리와 연산 비용 절감.

1. **양자화(4bit)**:

전체 모델 메모리 사용량을 최소화하고, GPU 없는 라즈베리파이에서도 실시간 추론 가능.

실험 결과, 제안된 경량화 파이프라인은 **10초 이내 응답**을 달성하였으며, 제한된 연산 자원 환경에서도 안정적인 모델 동작이 가능함을 확인하였다.

5.실험결과 ( 표도 넣어야 함)

 본 연구에서 제안한 토마토 스마트팜 특화 소형 언어모델(sLM)의 성능을 평가하기 위해 라즈베리파이 모델 \_\_\_(예: Raspberry Pi 4 Model B, 8GB RAM)\_\_\_를 사용하였다. 운영체제는 \_\_\_(예: Raspberry Pi OS 64-bit)\_\_\_이며, 모델 구현에는 Python \_\_\_(버전 \_\_\_)\_\_\_과 PyTorch \_\_\_(버전 \_\_\_)\_\_\_를 사용하였다. 또한 Transformers \_\_\_(버전 )및 NumPy \_\_\_(버전 ) 등의 라이브러리를 활용하였다.

모델 성능 평가는 세 가지 지표를 중심으로 진행하였다. 첫째, 정확도(Accuracy)는 센서 데이터 기반 질문에 대해 모델이 생성한 답변의 정확성을 평가하였으며, 평가 방식은 \_\_\_(예: 사람이 직접 채점하거나 정답 데이터셋과 비교)\_\_\_으로 수행하였다. 둘째, 응답 속도(Latency)는 모델 입력 후 출력까지 걸리는 시간을 측정하였으며, 목표는 10초 이내의 응답 달성이다. 셋째, 모델 크기(Model Size)는 경량화 전후의 파일 용량(MB)을 비교하여, 모델 압축 효과를 평가하였다. 경량화 단계는 원본 Gemma 1B 모델을 시작으로 1차 프루닝, LoRA기반 파인튜닝, 2차 프루닝, LoRA 반복, 양자화 순으로 진행되었다.

실험 결과, 각 경량화 단계에서 모델 크기와 응답 속도, 정확도는 다음과 같이 나타났다. 원본 Gemma 1B 모델의 크기는 \_\_\_\_ MB였으며, 평균 응답 시간은 \_\_\_\_초, 정확도는 \_\_\_\_%였다. 1차 프루닝을 적용한 후 모델 크기는 \_\_\_\_ MB로 감소하였고, 응답 속도는 \_\_\_\_초, 정확도는 \_\_\_\_%로 나타났다. RoLA 기반 파인튜닝 단계에서는 정확도 \_\_\_\_%를 달성하면서 모델 크기 \_\_\_\_ MB, 응답 시간 \_\_\_\_초를 기록하였다. 2차 프루닝과 RoLA 반복을 거친 후 모델 크기는 \_\_\_\_ MB, 응답 속도 \_\_\_\_초, 정확도 \_\_\_\_%로 나타났으며, 마지막으로 4비트 양자화를 적용한 최종 모델의 크기는 \_\_\_\_ MB로 줄어들고, 응답 시간은 \_\_\_\_초, 정확도는 \_\_\_\_%로 나타났다.

이 결과를 통해 제안한 경량화 파이프라인이 라즈베리파이와 같은 제한된 연산 자원 환경에서도 목표한 10초 이내 응답 속도를 달성할 수 있으며, 모델의 메모리 사용량을 크게 줄이면서 도메인 특화 성능을 유지할 수 있음을 확인하였다. 향후 실제 스마트팜 환경에 적용하여 다양한 센서 데이터와 사용자 질의에 대응할 수 있는 가능성을 제시한다.

5. 결론

 본 연구에서는 토마토 스마트팜 환경에서 센서 데이터를 활용하는 소형 언어모델(sLM)의 경량화 방법을 제안하였다. Gemma 1B 모델을 기반으로 구조적 프루닝, LoRA기반 파인튜닝, 반복적 프루닝, 양자화를 적용하여, 라즈베리파이와 같은 제한된 연산 자원 환경에서도 10초 이내 응답이 가능한 효율적인 모델을 구현하였다. 실험 결과, 제안한 경량화 파이프라인은 모델의 메모리 사용량과 연산량을 크게 줄이면서도 도메인 특화 성능을 유지할 수 있음을 확인하였다.

이로써 제한된 자원 환경에서 실시간으로 동작 가능한 스마트팜 특화 sLM의 가능성을 입증하였으며, 저비용 AI 보조 시스템 구축에 기여할 수 있음을 보여준다. 다만 본 연구에서는 전체 챗봇 시스템 구현보다는 언어모델 계층의 경량화에 초점을 맞추었기 때문에, 향후 연구에서는 데이터 처리 모듈 및 사용자 인터페이스와 통합된 종합 시스템 개발, 다양한 작물과 환경에 대한 도메인 확장, 추가적인 모델 최적화 기법 적용 등을 통해 보다 실용적인 스마트팜 AI 시스템으로 발전시킬 필요가 있다.

1. 개요

 - 현재 문제점 제시

 - 구현 시스템 기반 해결 가능성

1. 관련 연구
   1. 작물 재배 관련해서 llm 파인튜닝한 논문 찾고 소개한 후 하지만 이거는 고사양 환경에서만 쓸 수 있다. 우리는 경량화를 통해 이러한 한계를 극복했다.

AgriSentinel: Privacy-Enhanced Embedded-LLM Crop Disease Alerting System ←— 온디바이스 llm으로 작물 병해 대응 가이드 제공 연구, gpt 왈 튜닝 시 사용한 데이터가 1200개로 그리 많지 않다고 함, 온디바이스로 쓰기에 여전히 메모리/추론시간 측면에서 부담이 있을 가능성 있다고 함

1. 스마트팜이 이러저러하다는 논문 소개하고 스마트팜의 문제 설명한 후 우리 slm이 그런 문제를 해결할 수 있다. (농부가 토마토를 기르면서 의사결정을 할 때 도움을 주는 역할.)

1. 시스템 상세 설계
   1. 모델 선택하고 프루닝, lora, 양자화, GGUF 변환한 과정
   2. 파인튜닝할 때 어떤 데이터 썼는지 설명 (데이터 카테고리)

1. 시스템 구현 결과
   1. 경량화해서 이정도 리소스만 쓰는 모델을 만들었다

기존 baseline 모델과 우리가 만든 모델과의 표로 성능 비교 (추론 속도, 답변의 정확도, 메모리 사용량)

1. 라즈베리파이 저사양 환경에서 실제로 sLM 모델이 돌아가더라.

1. 결론
   1. 기능과 의의의 간략히 설명하고 향후 이러저러한 분야에도 적용 가능하다
2. 참고문헌

[1] <https://www.smartfarmkorea.net/>

[2] 이현병, 최인찬, 임동혁, 허정욱, 김태현, 이정민, 이정호, 박성진, 이태석, 백정현, “LLM과 RAG를 활용한 스마트팜 제어 및 관리 지원 시스템 설계 및 구현,” 한국콘텐츠학회논문지, 제24권, 제12호, pp.65-76, 2024.

[3] <https://developers.googleblog.com/ko/gemma-explained-whats-new-in-gemma-3/>

[4] AgriNex: Next-Gen Smart Agriculture with LLM-Integrated UAV-IoT Solutions

구조적 프루닝을 사용한 MobileViT 경량화