**토마토 스마트팜 특화 소형 언어모델의 효율적 경량화 연구**

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

**초록**

 본 연구는 토마토 스마트팜 환경에서 실시간 질의응답이 가능한 경량화 언어모델 기반 챗봇 시스템을 제안한다. 기존 대규모 언어모델(LLM)은 높은 연산 비용과 메모리 요구로 인해 농가의 온디바이스 환경에서 활용이 제한적이었다. 이를 해결하기 위해 Gemma 1B 모델을 기반으로 데이터 기반 구조적 프루닝, LoRA 기반 파인튜닝, 그리고 4비트 양자화를 순차적으로 적용한 3단계 경량화 파이프라인을 구축하여 소형 언어모델(sLM)을 구현하였다. 제안한 방법은 메모리 사용량과 연산량, 추론 시간을 크게 줄이면서도 도메인 특화 질의응답 성능을 유지하여, 저비용·고효율의 스마트팜 AI 보조 시스템 실현 가능성을 보여준다. 본 연구의 결과는 향후 스마트농업 인공지능 분야의 경량 언어모델 연구 및 온디바이스 대화형 시스템 응용에 활용될 수 있을 것으로 기대된다.

1.서론 

전 세계적으로 스마트팜 기술이 확산되고 있으며, 특히 토마토 재배는 온도, 습도, CO₂농도, 광량 등 환경 조건에 따라 수확량과 품질이 크게 달라 데이터 기반 관리가 중요하다. 최근 농업 분야에서는 생성형 AI를 활용한 챗봇 서비스가 등장하고 있으며, 국내에서도 농촌진흥청이 초거대 언어모델 기반의 '최신농업기술알리미' 앱을 통해 실시간 영농 상담 서비스를 제공하고 있다. 그러나 현장에서는 전문 지식 부족과 복잡한 데이터 해석 문제로 어려움을 겪고 있으며, 대규모 언어모델(LLM)은 연산 비용과 메모리 요구로 인해 농가 실용성이 낮다. 본 연구는 Gemma 1B 모델에 가중치 프루닝, LoRA 파인튜닝, 양자화를 적용하여 라즈베리파이 환경에서 10초 이내 응답 가능한 경량화 챗봇을 제안하며, 스마트팜 현장에서 활용 가능한 저비용·고효율 AI 시스템의 기반을 마련하고자 한다.

2. 관련연구  
 스마트팜 분야에서는 IoT 센서를 활용한 환경 모니터링과 작물 생육 예측 연구가 활발히 진행되고 있으며, 최근 대규모 언어모델(LLM)을 활용한 도메인 특화 AI 연구에서는 모델 경량화 기술의 중요성이 부각되고 있다. Hu et al. (2021)이 제안한 LoRA(Low-Rank Adaptation)는 저차원 적응 행렬 학습을 통해 훈련 파라미터를 10,000배 감소시키면서도 전체 파인튜닝과 동등한 성능을 달성했다. Sun et al. (2023)의 Wanda는 재훈련 없이 구조적 프루닝으로 효율적인 경량화를 실현했으며, Nagel et al. (2021)은 포스트 트레이닝 양자화로 저비트 환경에서도 높은 성능을 유지하는 방법론을 제시했다. 그러나 기존 연구는 고성능 GPU를 전제로 하여 현장 스마트팜 환경에 직접 적용하기 어려웠다. 본 연구는 이러한 경량화 기법들을 토마토 스마트팜 도메인에 통합 적용하여 라즈베리파이 환경에서 실시간 대화형 AI 구현 가능성을 제시한다."

3. 제안 방법

본 연구는 토마토 스마트팜 질의응답을 위한 온디바이스 경량 언어모델 파이프라인을 제안한다. 전체 과정은 데이터 기반 구조적 프루닝, LoRA 기반 미세조정, 4bit 양자화의 세 단계로 구성된다.

3.1. 데이터 수집 및 특화 포맷 정제

토마토 스마트팜 환경에서 농업인의 실제 질의 패턴을 반영한 6개 카테고리 질의응답 데이터셋을 수집하였다. 모든 데이터는 센서 모니터링 정보를 질문 앞에 명시하는 포맷으로 구조화하여, 모델이 환경 조건 기반 응답을 생성하도록 학습시킨다. 데이터는 계층적 분할로 각 카테고리 분포를 유지하며 프루닝셋(40%), 미세조정셋(40%), 검증셋(20%)으로 분할하였다.

3.2. 활성화 기반 FFN 뉴런 프루닝

모델 경량화를 위해 활성화 기반(Activation-based) 구조적 프루닝을 수행하며, Attention Head는 유지하고 FFN 뉴런에만 집중한다. 프루닝셋의 forward pass를 통해 각 FFN 뉴런의 활성화 통계를 측정하고, 이에 가중치를 부여하여 중요도 점수를 산출한다. 초기, 중간, 후기 레이어에 따라 중요도 가중치를 차등 적용하며, 레이어당 최대 제거 비율 및 최소 보존 뉴런 수 제약을 설정하여 성능 손실을 방지한다. 이를 통해 모델 파라미터를 약 10−30% 감소시킨다.

3.3. LoRA 기반 미세조정 및 모델 통합

프루닝 후 성능 보존 및 스마트팜 도메인 특화를 위해 LoRA 기법을 적용한다. Self-Attention 및 Feed-forward 계층의 주요 projection에 LoRA 모듈을 삽입하고 미세조정용 데이터셋으로 학습시킨다. 학습 완료 후 LoRA 가중치를 프루닝된 기저 모델과 병합하여 단일 모델로 통합한다.

3.4. 양자화 및 온디바이스 최적화

최종적으로 LoRA가 병합된 모델을 GGUF 포맷으로 변환하고 Q4\_K\_M 방식의 4bit 양자화를 적용한다. 이 혼합 정밀도 전략은 모델 크기를 약 75% 감소시키고 CPU 기반 온디바이스 환경에서 실시간 추론을 가능하게 한다.

3.5. 검증

경량화 모델의 성능은 최종 검증셋으로 텍스트 품질과 효율성을 평가한다. 텍스트 품질은 ROUGE-L과 BLEU 점수로 검증하며, 효율성은 평균 응답 시간, 메모리 사용량, 파라미터 수를 측정한다. 라즈베리파이 환경에서 10초 이내 응답을 목표로 GPU 클라우드 및 로컬 CPU 환경에서 추론 성능을 검증하여 배포 가능성을 입증한다. 각 단계별 성능 변화와 경량화 효과를 정량 분석하여 파이프라인의 유효성을 입증하였다.

4. 실험 결과

5. 결론

 본 연구에서는 토마토 스마트팜 환경에서 활용 가능한 소형 언어모델(sLM)의 효율적 경량화 방법을 제안하고 구현하였다. Gemma 1B 모델에 구조적 프루닝, LoRA 기반 파인튜닝, 4비트 양자화를 순차적으로 적용하는 경량화 파이프라인을 구축함으로써, 연산 자원이 제한된 라즈베리파이 CPU 환경에서도 10초 이내에 응답이 가능한 모델을 성공적으로 구현하였다.

실험 결과, 제안한 방법은 모델의 메모리 사용량과 연산량을 크게 줄이면서도, 토마토 스마트팜 도메인 특화 질의응답 성능을 유지함을 확인하였다. 이는 고성능 하드웨어 없이도 현장에서 실시간으로 동작하는 AI 보조 시스템 구현이 가능함을 의미하며, 대규모 LLM 적용 시 발생하는 높은 연산 비용과 자원 요구 문제를 효과적으로 완화하여 경제적 부담을 줄였다는 점에서 의의가 있다.

따라서 본 연구는 농가에 저비용·고효율 AI 시스템을 제공하여 운영 비용 절감과 생산성 향상에 기여할 수 있는 실용적 방향을 제시한다. 향후 연구에서는 제안 모델에 센서 데이터 처리 및 사용자 인터페이스를 통합한 종합 챗봇 시스템 개발, 다양한 작물 및 환경으로의 도메인 확장, 온디바이스 환경 최적화 추가 경량화 기법 적용을 통해 스마트팜 AI 솔루션의 실용성과 범용성을 더욱 향상시킬 예정이다.

참고문헌

[1] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[2] Sun, M., Liu, Z., Bair, A., & Kolter, J. Z. (2023). A Simple and Effective Pruning Approach for Large Language Models. arXiv preprint arXiv:2306.11695.

[3] Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M., & Blankevoort, T. (2021). A White Paper on Neural Network Quantization. arXiv preprint arXiv:2106.08295.