**토마토 스마트팜 특화 소형 언어모델의 효율적 경량화 연구**

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

초록쓸거임

본 연구에서는 XXX 시스템의 효율성을 평가하고, YYY 알고리즘을 기반으로 한 ZZZ 모델을 제안한다. 실험 결과, 제안 모델은 기존 방법 대비 성능 향상을 보였으며, 특히 AAA 지표에서 높은 정확도를 달성하였다. 또한, BBB 환경에서의 적용 가능성을 검토한 결과, 시스템 안정성 및 처리 속도 측면에서도 유의미한 개선이 확인되었다. 본 연구의 결과는 향후 CCC 분야 연구 및 DDD 응용에 활용될 수 있을 것으로 기대된다.

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 분야에서는 IoT 센서를 활용한 환경 모니터링과 작물 생육 예측 연구가 활발히 진행되고 있으며, 데이터 기반 의사결정과 자동화 시스템의 필요성이 강조되고 있다. 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 사전 훈련된 모델의 가중치를 고정한 채 저차원 적응 행렬만을 학습함으로써 훈련 가능한 파라미터를 10,000배 감소시키면서도 전체 파인튜닝과 동등한 성능을 달성할 수 있음을 입증했다. 모델 프루닝 분야에서 Sun et al. (2023)의 Wanda 방법은 재훈련 없이 가중치 크기와 활성화 값을 동시에 고려한 구조적 프루닝을 통해 대규모 언어모델의 효율적인 경량화를 실현했다. 신경망 양자화 분야에서는 Nagel et al. (2021)이 포스트 트레이닝 양자화와 양자화 인식 훈련 기법을 통해 저비트 가중치와 활성화를 유지하면서도 부동소수점 정확도에 근접한 성능을 달성하는 방법론을 제시했다.

그러나 기존 연구는 대부분 클라우드 환경이나 고성능 GPU를 전제로 하여, 현장 스마트팜 환경에 직접 적용하기에는 한계가 있었다. 본 연구는 이러한 경량화 접근법들을 토마토 스마트팜 도메인에 통합 적용하여 라즈베리­파이 CPU 환경에서 실시간 대화형 AI 구현 가능성을 제시한다.

3. 제안 방법

본 연구는 토마토 스마트팜 질의응답 시스템을 위한 온디바이스 경량 언어모델 파이프라인을 제안한다. 전체 과정은 데이터 기반 구조적 프루닝(Data-driven Structured Pruning), LoRA 기반 미세조정(Fine-tuning), 그리고 \*\*4bit 양자화(Quantization)\*\*의 세 단계로 구성된다.

1.데이터 기반 구조적 프루닝

토마토 Q&A 데이터셋(총 9,000문장)을 활용하여 모델의 forward pass를 수행하고, 각 Transformer layer의 attention head와 MLP neuron 단위로 평균 활성값을 기록한다.

수집된 활성도 분포를 분석하여, 평균 활성값이 낮은 하위 30%의 attention head와 MLP neuron을 제거한다.

이 방법은 단순 비율 프루닝이 아니라 도메인 특화 데이터 중심 접근으로, 모델 파라미터와 FLOPs를 약 30% 감소시키면서도 성능 손실을 최소화한다.

2.LoRA 기반 미세조정

프루닝으로 인한 표현력 손실을 보완하기 위해 LoRA 모듈을 적용한다.

Self-Attention 및 Feed-forward 계층의 주요 projection(q, k, v, o, up, down, gate)에 LoRA 모듈을 삽입하고, rank=16, α=32, dropout=0.1로 설정한다.

토마토 질의응답 데이터셋으로 3 epoch 학습하며, 배치 크기는 2, gradient accumulation step은 8로 설정한다.

학습 후 LoRA 모듈만 저장하고, 이후 기저 모델(Base Model)과 병합하여 하나의 완전한 모델 형태로 통합한다.

3.양자화 및 온디바이스 최적화

병합된 모델을 llama.cpp 변환 스크립트를 이용하여 GGUF 포맷으로 변환하고, \*\*4bit 양자화(Q4\_K\_M)\*\*를 적용한다.

이를 통해 모델 크기를 약 75% 감소시키고, CPU 환경에서도 실시간 응답이 가능하도록 최적화한다.

4. 실험결과

본 연구의 실험은 Gemma 1B 모델을 기반으로 진행되었으며, 토마토 스마트팜 질의응답 데이터셋(총 9,000문장) 중 982개를 검증용으로 사용하였다. 모델은 라즈베리파이 5 (CPU) 및 Mac M2 (GPU) 환경에서 테스트되었으며, 각 단계별 모델(기저 모델, 프루닝 모델, LoRA 적용 모델, 최종 4bit 양자화 모델)의 성능과 효율을 비교하였다.

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.

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. (19,319회 인용)

[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. (865회 인용)

[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. (871회 인용)

3. 방법론 및 제안방법

본 연구는 토마토 스마트팜 환경에 적합한 온디바이스 소형 언어모델(sLM)을 구현하기 위해, 데이터 기반의 구조적 프루닝(Data-driven Structured Pruning), LoRA(Low-Rank Adaptation), 그리고 양자화(Quantization)를 결합한 효율적인 경량화 파이프라인을 제안합니다. 실험에는 Gemma 1B 모델을 사용하였으며, 라즈베리파이 CPU 환경에서도 실시간 응답이 가능하도록 다음과 같은 단계를 적용했습니다.

본 연구의 pruning 과정은 데이터 기반 활성도(activation) 분석을 통해 수행되었다.

먼저 토마토 스마트팜 질의응답 데이터셋(총 9,000개 문장)을 이용하여 모델의 forward pass를 수행하고, 각 Transformer layer의 attention head와 MLP neuron 단위로 평균 활성값을 기록하였다. 수집된 활성도 분포를 분석하여, attention head는 평균 activation이 낮은 하위 30%를, MLP neuron은 평균 활성도 및 L1-norm이 낮은 하위 30%를 pruning 대상으로 선정하였다. 이 방법은 단순히 임의 비율로 제거하는 heuristic pruning이 아니라, 실제 도메인 데이터의 통계적 활성 패턴에 기반한 데이터 중심(Data-driven) structured pruning이다. 따라서 pruning 과정은 모델의 실사용 환경(토마토 QnA 질의응답)에 특화된 불필요 파라미터를 효과적으로 제거하며, 모델의 파라미터 수와 FLOPs를 약 30% 이상 감소시키면서도 응답 정확도의 손실을 최소화하였다.

프루닝으로 인해 모델의 일부 attention head 및 MLP neuron이 제거되면, 특정 표현 능력의 손실이 발생한다.

이를 완화하기 위해 본 연구에서는 **LoRA (Low-Rank Adaptation)** 기반 파인튜닝을 적용하였다.

LoRA는 기존 모델의 가중치를 직접 업데이트하지 않고, 각 선형변환(Linear Projection)에 대해 저랭크(rank-r) 행렬을 추가하여 학습하는 기법이다.

즉, 파라미터 효율적 튜닝(Parameter-Efficient Fine-Tuning, PEFT)의 일종으로, 원본 모델의 파라미터는 고정된 상태에서 LoRA 모듈의 파라미터만 학습된다.

구체적으로, 본 연구에서는 프루닝된 모델의 self-attention 및 feed-forward layer 내 주요 projection 연산(q\_proj, k\_proj, v\_proj, o\_proj, up\_proj, down\_proj, gate\_proj)에 LoRA 모듈을 삽입하였다.

랭크(r)는 16, scaling factor(α)는 32로 설정하였으며, dropout 비율은 0.1로 적용하였다.

이 설정은 모델 크기 증가를 최소화하면서도 pruning으로 인한 정보 손실을 보완하기 위한 경험적 최적값이다.

파인튜닝에는 \*\*토마토 질의응답 데이터셋(총 9,000개 문장)\*\*을 사용하였고, 각 데이터는 “질문(instruction) + 답변(output)” 형태로 구성하였다.

모델은 3 epoch 동안 학습되었으며, 배치 크기는 2, gradient accumulation step은 8로 설정하였다.

학습 후에는 LoRA 어댑터 가중치만 별도로 저장하여, 원본 프루닝 모델에 모듈형태로 결합 가능한 형태로 관리하였다.

결과적으로 LoRA 미세조정은 프루닝으로 인한 정확도 손실을 효과적으로 회복시켰으며, 모델의 추론 속도 및 경량성은 그대로 유지되었다.

LoRA 기반 미세조정이 완료된 후, 본 연구에서는 LoRA 어댑터를 \*\*기저 모델(Base Model)\*\*과 병합하여 하나의 완전한 모델(merged full model)을 생성하였다.

이 과정은 PeftModel.merge\_and\_unload() 함수를 통해 수행되며, LoRA 모듈의 저랭크 가중치가 원본 모델의 projection 행렬에 직접 통합된다.

병합된 모델은 Hugging Face 호환 형식으로 저장되며, 이후 추론 시 별도의 어댑터 로드 과정이 필요하지 않다.

이후 경량 디바이스 환경에서의 추론 효율을 극대화하기 위해, 모델을 \*\*GGUF 포맷(GerGANOV’s Unified Format)\*\*으로 변환하였다.

변환 과정은 llama.cpp 프로젝트의 convert\_hf\_to\_gguf.py 스크립트를 활용하여 수행되었으며, 최종 모델 가중치는 f32 형식으로 내보낸 뒤 \*\*4bit 양자화(Q4\_K\_M)\*\*를 적용하였다.

4bit 양자화는 각 weight tensor를 16분의 1 크기로 압축하면서, 주요 통계적 분포를 유지하도록 설계된 mixed-precision quantization 방식이다.

이로 인해 모델의 전체 크기가 약 75% 이상 감소하였고, 추론 속도는 GPU 및 CPU 환경 모두에서 약 2~3배 향상되었다.

특히 GGUF 포맷은 온디바이스 환경에서의 실행을 지원하여, 모델을 로컬 장치(Mac M-series, Jetson 등)에서도 원활히 동작시킬 수 있다.

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.