무선 로봇 작업 계획을 위한 불확실성-인식 기회 기반 하이브리드 언어 모델

박제영, 임연섭*, 김성륜*
워털루대학교, *연세대학교
j388park@uwaterloo.ca., *{yeonsub415, slkim}@yonsei.ac.kr

Uncertainty-Aware Opportunistic Hybrid Language Model for Wireless Robot Task Planning

Jeyoung Park, YeonSub Lim*, Seong-Lyun Kim*

요 약

최근 대규모 언어 모델(LLM)이 로봇 시스템에 통합되면서 로봇이 자연어를 수신하고 해석할 수 있게 하여 로봇 작업 과정이 향상되었다. 그러나 LLM을 이용한 로봇 작업 과정에서 지연 시간과 전력 소비를 증가시키는 상당한 메모리와 계산 요구를 필요로 한다. 반면, 온디바이스 언어 모델(SLM)은 종종 정확도가 감소하고 의미적 용량이 제한된다. 본 논문에서는 U-HLM(불확실성-인식 기회 기반 하이브리드 언어 모델)을 통해, LLM의 문제를 완화하는 동시에 SLM의 한계를 해결한다. 실제 무선 네트워크 환경에서 베이스라인 방법과 비교하여 성능을 평가하기 위해 U-HLM을 실시간 무선 Wi-Fi 테스트베드에 배포하고, 불확실성 계산을 선택적으로 건너뛰어 U-HLM의 지연 시간을 더욱 최소화하는 방법을 제안한다. U-HLM을 사용하여 높은 정확도와 높은 처리량 생성이 가능하다는 것을 실험을 통해 확인하였다.

Ⅰ. 서 론

최근 복잡한 로봇 작업을 수행에 대규모 언어 모델(LLM)을 활용한 로봇 작업 계획에 관한 연구가 진행되고 있다. [1] 하지만 이러한 발전에도 불구하고, 원격 추론과 통신 지연 등의 제한 사항으로 실시간 요구사항을 만족하기 어렵다. [2] 이를 해결하기 위해, 온디바이스 언어 모델(SLM)과 원격 LLM을 결합한 불확실성-인식 기회 기반 하이브리드 언어 모델(U-HLM)을 제안한다. [3]

선행 연구는 시뮬레이션된 무선 조건을 사용하여 U-HLM을 평가하고 모든 토큰에 대해 불확실성을 계산하여 불확실성 추정 자체의 오버헤드를 무시했다.[3] 이는 실제 무선 Wi-Fi 환경의 가변성을 포착하고 불확실성 계산의 오버헤드를 고려하는데 한계가 있고, 로봇 작업 실행을 위한 계획 파이프라인에 U-HLM을 통합하는 방법을 보여주지 않는다.

본 연구에서는 실시간 Wi-Fi 테스트베드에서 U-HLM을 배포하여 최적 불확실성 임계값을 도출하고 지연 요소를 세분화해 벤치마킹하고, 조건부 불확실성 평가를 도입해 추가 지연을 줄이며, 다양한 실제 로봇 작업 시나리오에서 U-HLM을 통합·검증한다. 실험 결과, 네트워크 상태에 따라 U-HLM은 빠른 실행 속도와 더 높은 정확도를 얻었으며, 선택적 불확실성 계산으로 불필요한 오버헤드를 최소화하여 LLM을 사용하면서 나타나는 지연과 신뢰성 문제를 극복하는 실용적인 대안임을 확인하였다.

Ⅱ. 시스템 모델

불확실성-인식 기회 기반 하이브리드 언어 모델(U-HLM)은 자원이 제한된 로컬 디바이스에 배포된 소형 언어 모델(SLM)과 고성능 서버에 존재하는 대형 언어모델(LLM)의 두가지 요소로 구성된다.[3] U-HLM의 핵심 아이디어는 로컬 디바이스에 배포된 SLM이 먼저 초안 토큰을 제안하고,자체 불확실성을 계산하는 것이다. 불확실성은 모델이 자체 출력에 대해 평가한 신뢰도이다. 불확실성이 미리 정의된 임계값을 초과하면, 토큰은 수락 또는 거부를 위해 LLM으로 전송된다.

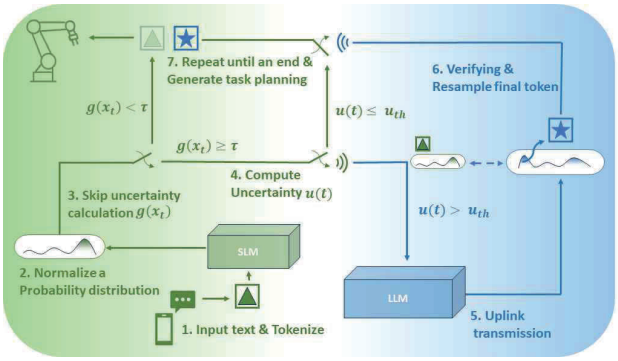

그림 1. 불확실성-인식 기회 기반 하이브리드 언어 모델 구조

Ⅲ. 베이지안 기반 선택적 불확실성 계산

이 절에서는 토큰별 불확실성 추정의 과도한 계산 오버헤드를 줄이기 위해 베이지안 결정 이론에 기반한 선택적 불확실성 계산 기법을 제안한다. 기존 방식은 온도 섭동(temperature perturbation)마다 Vocabulary size V만큼의 소프트맥스 연산을 K번 수행하여, 토큰당 O(KV)의 막대한 연산 비용을 야기한다. 이를 개선하기 위해 먼저 순차적 소프트맥스 호출로 각각의 실행 지연 시간을 분리 측정함으로써, 병렬 처리 시 발생하는 메모리 대역폭 경쟁과 스케줄링 오버헤드를 배제한 명확한 기준선을 확보한다.[4] 다음으로, 과거 k회의 불확실성 추정치 $u_{t-1}, \ldots, u_{t-k}$, 해당 시점의 최상위 토큰 확률 $p_{t-1}^{\max}, \ldots, p_{t-k}^{\max}$, 그리고 토큰 인코딩 ϕ를 하나의 고정 길이 벡터 x_t로 결합한다. 이 벡터를 입력으로 하는 로지스틱 회귀 분류기를 교차 엔트로피 손실로 학습하여, $g(x_t) \approx P(U_t > \theta | x_t)$를 예측하도록 함으로써 실제 불확실성이 임계값 θ를 초과할 확률을 사전에 평가할 수 있다. 마지막으로, 불확실성 계산에 드는 시간 비용 C_u와, 계산을 건너뛰었다가 잘못된 결정을 복구하는 데 소요되는 비용 C_e 를 정의한다. 베이지안 최소 위험 규칙에 따라 임계값 $\tau = \frac{C_u}{C_e}$를 도출하고, 분류기 예측값

$g(x_t)$이 τ보다 작을 때만 불확실성 계산을 건너뛰도록 함으로써, 불필요한 연산을 크게 줄이면서도 정확도 손실을 최소화한다.[5]

IV. 실험 결과

1) 지연 시간 감소를 위한 선택적 불확실성 계산 스킵

2500개의 토큰 생성 단계에서 분류기는 9.1%(227개)만큼 불확실성 계산을 건너뛰도록 선택했으며, 이 중 오류 스킵은 8.8%로 전체 오류율은 1% 미만이었다. 이를 통해 전체 불확실성 평가의 10% 가량을 제거하면서도 성능 저하를 거의 발생시키지 않음을 확인했다.

2) 유틸리티 최대화를 통한 임계값 선택

0부터 1까지 0.05 단위로 임계값을 조정하며 100개의 샘플을 테스트한 결과, 유틸리티 함수(신뢰도 지표 가중합 및 처리량 패널티) 최대화 지점은 u_{th}=0.15로 나타났다. 이 값에서 F1 점수가 최고치를 기록하면서도 토큰 처리량은 최대의 80.9%를 유지하여, 로봇 작업 계획에 적합한 최적 임계값임을 확인했다.

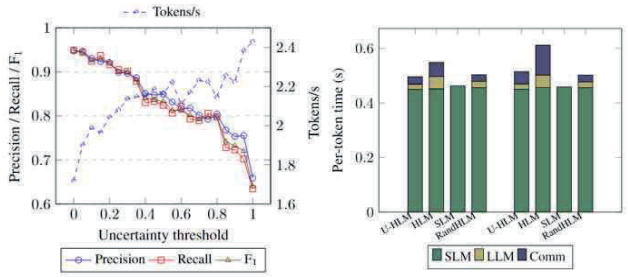

그림 2. Threshold senstibity, precision, recall, F1 vs 토큰처리량(좌) 네가지 방법의 토큰당 지연 시간(강한 vs 약한) (우)

3) 다양한 네트워크 조건에서의 기준선 간 비교 지연 시간 분석

두가지 실내 Wi-Fi 시나리오(강한/약한 커버리지)에서 SLM, Rand-HLM, HLM과 비교한 결과, u_{th}=0.15인 U-HLM은 열악 조건에서도 토큰당 지연 시간이 SLM 대비 12%, Rand-HLM 대비 2.6%만 증가했으나 F1은 각각 0.663에서 0.929, 0.703에서 0.929로 30% 이상 향상되었다. HLM 대비 지연은 19% 줄면서도 F1 손실은 2%미만에 그쳐, 통신 품질 악화에도 일관되게 낮은 지연과 높은 신뢰도를 유지함을 보였다. 아래의 식은 Rijsbergen, C. J. V. Information Retrieval에서 발췌하였다.

$$precision = TP/(TP+FP), recall = TP/(TP+FN) \quad (1)$$

$$F_1 = 2 \cdot \frac{precision \times recall}{precision + recall} \quad (2)$$

4) 실제 로봇 시스템 통합 평가

Doosan A0912s 로봇과 커피 서비스 시나리오에서, 쉬움/중간/어려움 세 난이도 모두에서 U-HLM은 전송률을 50-66% 절감하고 추론 시간을 10-20% 단축했다. 계획 성공률은 중간 난이도에서 0.70에서 0.65로 소폭 하락했으나, 쉬움과 어려움 수준에서는 HLM과 동등하게 유지되었다. 이를 통해, U-HLM은 실제 로봇 환경에서도 네트워크 트래픽과 지연을 크게 줄이면서, 높은 작업계획 신뢰성과 안전성을 보장하는 실용적 대안임을 입증한다.

III. 결론

본 논문에서는 계산 오버헤드를 더욱 줄이기 위한 선택적 불확실성 계산 스킵을 제안하고, 실제 무선 네트워크에서 기준 방법에 대한 U-HLM의 상대적 성능을 검증하며, 로봇 공학에서 U-HLM의 실제 구현을 보여주었다. 본 실험을 통해 이러한 선택적 스킵이 무시할 수 있는 수준의 성공률 감소만으로 종단간 지연 시간과 업링크 트래픽을 모두 절반으로 줄인다는 것을 확인했으며, 이는 로봇 플랫폼에서 U-HLM의 실시간 구현 가능성을 보여주고 적응형 임계값 설정과 더 넓은 작업 시나리오에 대한 후속 연구를 촉진한다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00347, 6G 통신을 위한 Post MAC, No.2022-0-00420,6G 종단간 초정밀 네트워킹을 위한 핵심기술 개발)

참 고 문 헌

[1] Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., et al., Foundation models in robotics: Applications, challenges, and the future, 2023 (https://arxiv.org/abs/2312.07843)

[2] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., F et al., Do as i can, not as i say: Grounding language in robotic affordances, 2022 (https://dl.acm.org/doi/10.1145/3442188.3445922)

[3] Oh, S., Kim, J., Park, J., Ko, S.-W., Quek, T. Q. S., and Kim, S.-L. Uncertainty-aware hybrid inference with on-device small and remote large language models, 2025 (https://arxiv.org/abs/2412.12687)

[4] Yang, M., Otterness, N., Amert, T., Bakita, J., Anderson, J. H., and Smith, F. D. Avoiding Pitfalls when Using NVIDIA GPUs for Real-Time Tasks in Autonomous Systems. In Altmeyer, S. (ed.), 30th Euromicro Conference on Real-Time Systems (ECRTS 2018), volume 106 of Leibniz International Proceedings in Informatics (LIPIcs), pp. 20:1 - 20:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl - Leibniz-Zentrum f ¨ur Informatik. ISBN 978-3-95977-075-0. doi: 10.4230/LIPIcs.ECRTS.2018.20.(https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ECRTS.2018.20)

[5] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification (2nd Edition). Wiley-Interscience, USA, 2000. ISBN 0471056693.