사회 • 사건사고

서울대 RTOSLab, AI 모델 최적화 국제대회서 2위

국산 NPU 하드웨어 구조 분석해

소프트웨어 최적화로 연산 효율↑

초당 821.1토큰 처리 성능 기록

정유나 기자

입력 2026-11-02 15:02



▲ 서울대 전기정보공학부 대학원생들이 1일(현지시간) 그리스 아테네에서 열린 IEEE/ACM 국제 마이크로아키텍처 심포지엄(MICRO 2026)의 'NPU 모델 최적화 경진대회'에서 2위를 차지했다. 사진 제공=서울대

서울대 연구팀이 인공지능(AI) 모델을 최적화하는 국제 경진대회에서 2위를 차지했다.

2일 서울대에 따르면 전기정보공학부 실시간 운영체제 연구실(RTOSLab)에서 석사과정을 밟고 있는 윤재민·박장운 씨가 IEEE/ACM 국제 마이크로아키텍처 심포지엄(MICRO 2026)에서 열린 'NPU 모델 최적화 경진대회'에서 2위에 올랐다.

이번 대회는 AI 반도체 기업 퓨리오사AI의 신경망처리장치(NPU) '레니게이드(RNGD)'에서 구글의 멀티모달 AI 모델 'Gemma 4 12B'의 처리 속도를 높이는 방식으로 진행됐다. 1라운드에서는 100개 이상의 참가팀이 AI 모델의 핵심 연산 세 가지를 최적화하는 과제를 수행했다. 이를 통과한 64개 팀은 2라운드에서 AI 모델 전체의 실제 서비스 처리 성능을 개선했다.

서울대 RTOSLab 팀은 여러 사용자의 요청을 한꺼번에 처리하는 추론 엔진을 새롭게 설계했다. NPU가 작업을 수행하는 동안 다음 작업을 미리 준비하도록 해 대기 시간을 최소화하고, 정확도를 유지하는 범위에서 연산 정밀도를 낮춰 처리 효율을 높였다. 이를 통해 추론 엔진의 성능을 5배 이상 끌어올렸으며 최종적으로 초당 821.1토큰의 처리 속도를 기록했다.

홍성수 서울대 전기정보공학부 교수는 "AI 반도체의 서비스 성능은 칩뿐 아니라 소프트웨어에 크게 좌우된다"며 "국산 AI 반도체의 하드웨어 구조를 직접 분석하고 이에 맞춰 소프트웨어를 최적화해 성능을 높였다는 점에서 의미가 있다"고 강조했다.

ⓒ 서울경제신문, 무단 전재-재배포, AI 학습 및 활용 금지