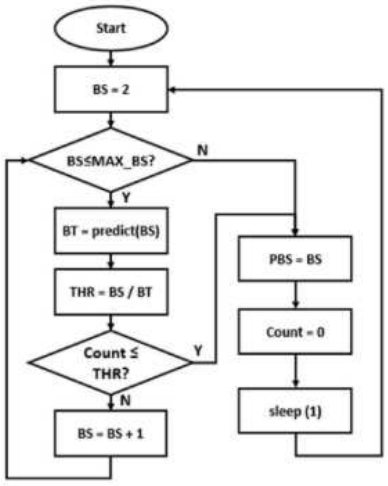
최적의 배치 크기를 결정하는 컴퓨팅 시스템

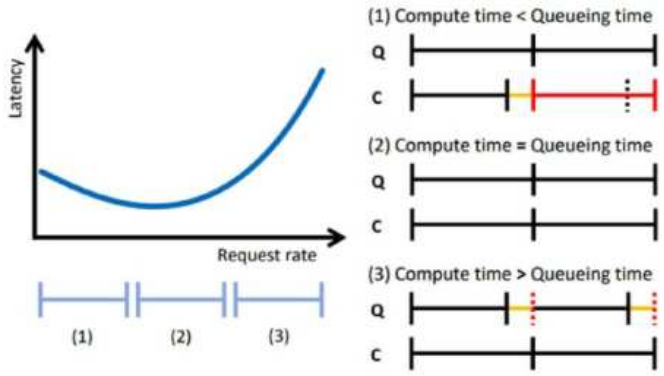
연구개발자: 소프트웨어학과 서의성 교수

I 기술 개요

01 기술 요약



[적응형 배치 크기 결정 방법에 대한 프로파일 데이터의 플로우 차트]



[추론 요청량에 대해 고정된 배치 크기로 요청을 처리하는 경우 나타나는 3가지 상황]

- 본 기술은 GPU 기반 기계학습 추론 서비스에서 실시간 요청 부하를 고려하여 배치 크기를 동적으로 결정하는 방법 및 장치에 관한 것으로, 이를 통해 지연 시간과 에너지 효율성이라는 상충 목표를 동시에 최적화하여 고성능·저비용 AI 인프라 구축을 가능하게 함

02 지식재산권 현황

No	발명의 명칭	출원번호	출원일
1	학습 클러스터에서 자원 사용률을 극대화하기 위한 유휴 cpu 자원 기반 심층 신경망 추론 방법	2023-0194699	2023.12.28
2	기계학습 추론 서버, 및 기계학습 추론 서비스에서의 적응형 배치 크기 결정 방법	2023-0193532	2023.12.27
3	기계학습 추론을 위한 gpu 클럭 조절 방법 및 장치	2022-0005427	2022.01.13

최적의 배치 크기를 결정하는 컴퓨팅 시스템

03 기술의 우수성

- **상충 목표 동시 최적화**

-지연 시간(Latency) 최소화와 에너지 효율성(Request per Joule) 최대화라는 상호 대립되는 두 목표를 동시에 충족시키는 독자적인 최적화 로직을 제공함

- **압도적인 성능 개선 효과**

-최대 배치 크기(Max Batch Size) 방식 대비 평균 지연 시간을 약 14배 빠르게 단축시키며, 기존 최적화 방식인 Ad-hoc Batching 대비 에너지 효율을 약 5% 향상시킴

- **실시간 부하 대응력**

-실제 워크로드와 같이 요청량이 크게 변동하는 환경에서도 실시간 대기열 크기에 기반하여 배치 크기를 동적으로 조정함으로써, 성능 저하 없이 안정적인 서비스 품질(QoS)을 유지함

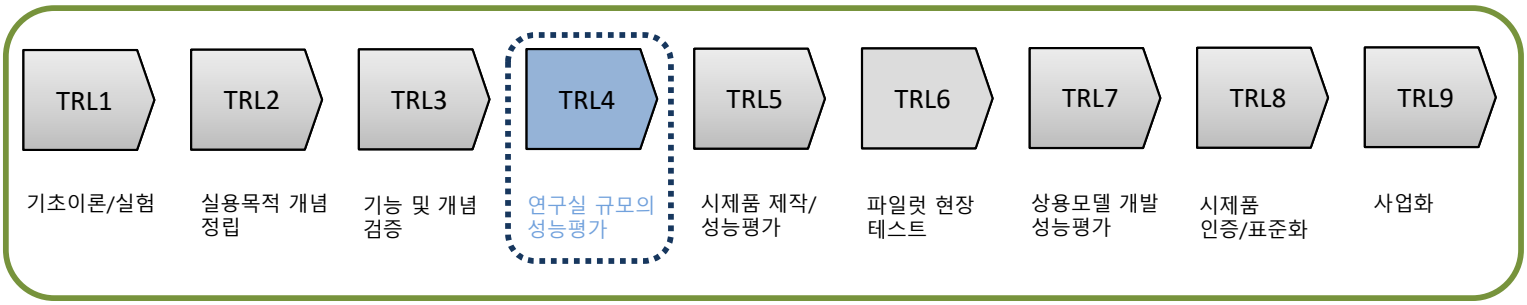
- **범용적인 적용 가능성**

-이미지 처리 모델(VGG19)과 언어 처리 모델(BERT-base)에 모두 적용되어 성능 개선 효과가 입증되었으며, 다양한 AI 모델(CV, NLP, LLM)의 추론 환경에 즉시 적용 가능

- **AI 인프라 운영 비용 절감**

-에너지 효율성 증가는 곧 GPU 서버 운영에 필요한 전력 비용을 직접적으로 절감시켜, 클라우드 서비스 제공자(CSP) 및 AI SaaS 기업의 수익성을 대폭 개선시킴

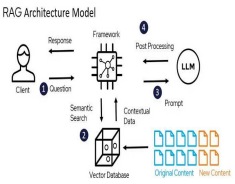
04 기술 개발 완성도

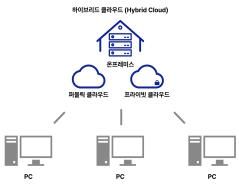


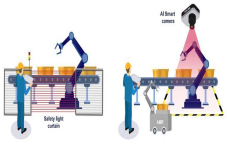
최적의 배치 크기를 결정하는 컴퓨팅 시스템

II 기술 동향

01 기술응용분야







[자율주행 및 로보틱스]

저전력/초저지연으로 센서 데이터 및 비전 모델 실시간 추론

[LLM 서비스]

폭발적인 요청에도 고효율 및 실시간 응답 지연 최소화

[클라우드 AI 인프라]

대규모 GPU 팜의 처리량 극대화 및 운영 전력 비용을 절감하는 솔루션

[산업/제조 비전 AI]

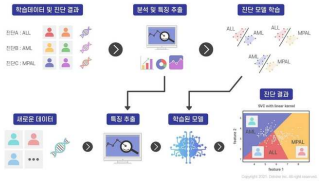
실시간으로 불량품을 감지하는 비전 AI의 추론 속도 향상

[실감 미디어/메타버스 플랫폼]

AI 객체 상호 작용 및 실시간 렌더링 시 고속의 AI 추론 지원

02 기술 동향

[2016~2018]

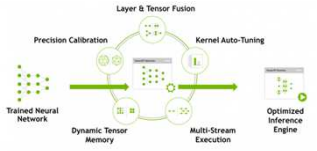


모델 정확도 우선 단계

[2019~2021]

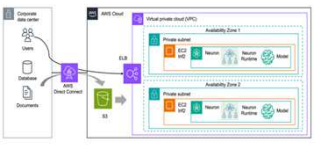
클라우드 확산과 지연 시간 경쟁

[2022~2023]



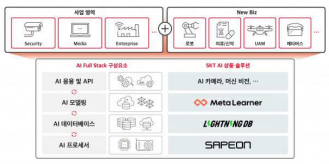
추론 가속화 및 효율성 부상

[2024~현재]



LLM 시대의 비용/전력 최적화

[향후 전망]



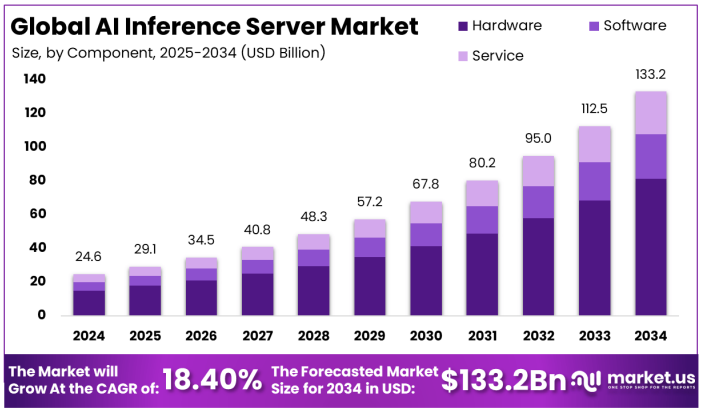
HW/SW 통합 최적화

AI 기술은 단순 정확도 경쟁을 넘어, 클라우드 및 엣지 환경에서 추론 서비스의 '경제성'과 '속도'를 동시에 확보하는 방향으로 진화했고, 현재는 대규모 LLM 및 Vision AI의 운영 비용 증가로 인해, 지연 시간을 유지하면서 에너지 효율을 극대화하는 본 기술과 같은 동적 최적화 솔루션이 시장의 핵심 요구 사항으로 부상하고 있음

최적의 배치 크기를 결정하는 컴퓨팅 시스템

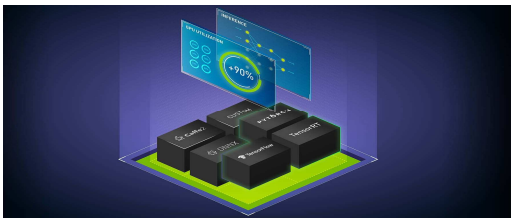
III 시장 동향

01 시장규모



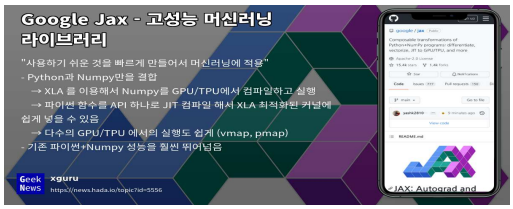
- 글로벌 AI 추론 시장은 2024년 246억 달러에서 2034년 약 1,332억 달러로 성장할 것으로 예상되며, 2025년부터 2034년까지 CAGR 18.40%로 성장할 것으로 전망됨

02 주요 시장 참여자



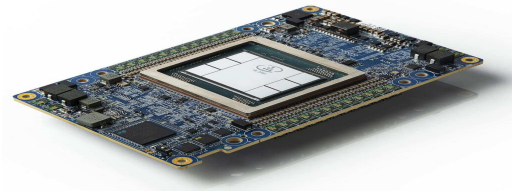
[NVIDIA社 TensorRT/Triton Inference Server 기술]

- GPU 기반 추론 최적화 엔진 및 서비스 플랫폼. 배치 처리 기능을 포함하여 성능을 극대화함



[Google社 TPU / JAX 기술]

- 자체 개발한 AI 가속기와 프레임워크를 통해 클라우드 환경에서 추론 효율성 극대화함



[Intel社 OpenVINO / Habana Gaudi 제품]

- CPU 및 통합 GPU, Gaudi 칩을 위한 추론 최적화 툴킷으로, Edge 및 Data Center 환경을 지원함

기술 이전 상담 및 문의

산학사업팀 담당자 김민경 / 031-290-5087 / mikkypat@skku.edu