인적 사항

이름 : 차주형 (JooHyoung Cha)
생년월일 : 1999.11.30
E-mail : chacha@udon.party
전화번호 : 010 - 7440 - 1754

포트폴리오 / 지식공유 활동

github.com/Piorosen
gitlab.com/Piorosen
blog.udon.party

클릭 시 해당 링크로 이동합니다

자기 소개

책임이 주어지거나 하고픈 일이 생긴다면, 다른 분야이더라도 며칠간 몰두하여 전력으로 탐구해내곤 합니다.
정확한 해답을 얻어내기 위해 해외의 개발자에게 이슈나 메일을 보내기도 하였고, 종종 해당 프로젝트에 기여[Android][Arm]하기도 했습니다.

딥 러닝 추론 가속과 리눅스 커널, 엔지니어링, 인프라 관리 및 개선한 것이 저의 대표적인 기술입니다.
임베디드 시스템에서 하드웨어의 자원을 최적화하여 TFLite의 성능을 최대 3.6 배 개선하기도 하였습니다.

효율적인 추론 구조를 설계하고, 메모리와 연산 구조에 따른 성능 개선에 대해 연구하여 기여하고자 합니다.
이를 위해 객관적인 자료 수집과 수용적인 태도로 최적화된 결과를 도출해내겠습니다.

학력

2018.03 - 2024.02 동의대학교 응용소프트웨어공학과 졸업예정 | 3.88 / 4.5
2015.03 - 2018.02 부산 동성고등학교 졸업

특허

10-2022-0125883	클라우드 노트북	출원
10-2281266	영상 내 자막 키워드 추출 및 순위 산정 시스템 및 방법 [url]	등록
2022 - 2023	한국저작권위원회 소프트웨어 저작권 등록 (8건)	등록

저널 (5 건)

JIEIE, Vol.60(7), 40-49, July. 2023 [pdf]
Profile-based Optimization for Deep Learning on Heterogeneous Multi-core CPUs | 1저자

JKIICE, Vol.24(2), 186-191, Feb. 2020 [pdf]
An Optimization Method of Spatial Placement for Effective Vehicle Loading | 1저자

JKIICE, Vol.23(8), 896-902, Aug. 2019 [pdf]
An Effective Method for Generating Images Using Genetic Algorithm | 1저자

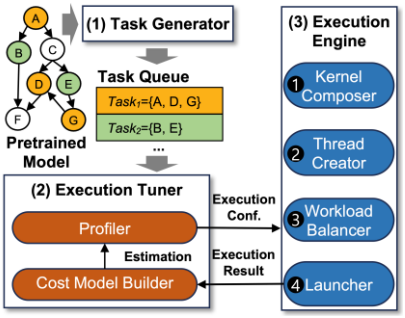
해외 컨퍼런스 (2 건)

Design Automation Conference 2024
(심사중) ACLTuner: A Profiling-Driven Fast Tuning to Optimized Deep Learning Inference | 2저자

ML for Systems at NeurIPS 2023 [pdf]
12.16 ACLTuner: A Profiling-Driven Fast Tuning to Optimized Deep Learning Inference | 2저자

ICFICE 2019, Vol.11(1) (355-358) [pdf]
06.25 – 06.27 An Effective Method for Generating Color Images Using Genetic Algorithm | 1저자

프로젝트 : 프로파일 기반으로 타겟 하드웨어에 최적화된 딥 러닝 추론 성능 개선



내용

TFLite에 대한 이기종 멀티 코어에서의 문제점을
코어마다 가장 적합한 작업 구성으로 프로파일링 수행

기관 : ETRI

기간 : 22.07.01 – 23.12.16

기술 :
Arm CPU, Mali GPU, Android , C/C++, TFLite, Container

업무

- 딥 러닝 추론 가속 라이브러리(TFLite, XNNPack, ComputeLibrary)의 소스코드 분석

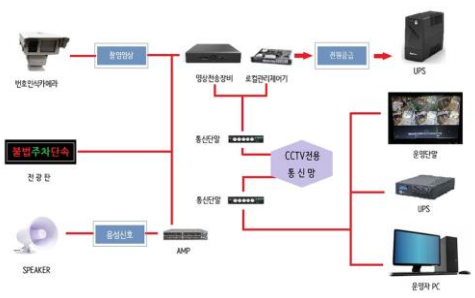
성과

- 이기종 멀티 코어에서는 성능 병목 현상에 대한 가능성 제기 및 개념 증명을 통한 문제점 발견
- 이기종 멀티코어를 대상으로 최적의 연산 작업 분배와 커널 탐색을 위한 프로파일러 구현 및 성능 데이터 수집
- 딥 러닝 최적화 도구인 AutoTVM, Ansor에 비해 최대 95%의 최적화 시간을 단축
- 1~2시간 내에 최적 결과 도출 가능
- Arm CPU와 Mali GPU 에 대해 최대 3.6배에서 2.8배 개선함

[1] JIEIE, Vol.60(7), 40-49, July. 2023 (**차주형**, 권용인, 이제민)
"Profile-based Optimization for Deep Learning on Heterogeneous Multi-core CPUs"

[2] ML for Systems at NeurIPS. Dec. 2023. (권용인, **차주형**, 이주빈, 유미선, 박제만, 이제민)
"ACLTuner: A Profiling-Driven Fast Tuning to Optimized Deep Learning Inference"

프로젝트 : 고정형태의 지능형 불법 주정차 단속 소프트웨어 개발



내용

도로상에 불법 주정차한 차량에 대해 영상검지 기법을 적용하여
자동 단속 및 고지서 발급하는 시스템 개발

기관 : 삼성 G2B (주) **인원** : 4명

기간 : 23.01.01 – 23.06.30

기술 :
TensorRT, Kafka, NVR, GStreamer, mariaDB, Container

업무

- 불법 주정차 단속을 위해 딥 러닝 기반의 객체 탐지 알고리즘 비교 및 선정
- RTSP 영상의 녹화와 데이터 처리를 위해 GStreamer 기반의 파이프라인 구성 및 NVR 시스템과 연동
- Kafka의 메시지 큐를 활용하여 불안정한 통신 환경에서도 단속 정보 및 환경 정보 동기화

성과

- 호모그래피 변환과 차량 탐지를 결합하여 주정차 영역에 있는 정적 차량을 탐지하는 모듈 개발
- 차량 탐지 모듈을 통해 차량 번호판이 2개 이상 탐지될 때 단속을 못하던 문제 개선함
- 메시지 큐의 도입 전/후로 YoloV3 모델과 Jetson AGX Xavier 보드에서 추론 처리량 성능이 약 1.7배 개선
- 고지서 발급을 위해 주변 사람과 주변 차량에 대한 자동 블러와 유사 차량 번호 추천 시스템 개발
- (진행중) GS 인증 및 조달등록

프로젝트 : 컨테이너와 딥 러닝 기반으로 인프라의 자원과 비용 최적화

내용

효율적인 컴퓨팅 자원 활용하기 위해 기존의 가상머신 기반에서 컨테이너 기반으로, NPU 기반의 영상처리로 수행하여 통신 비용 최적화

수업 : 캡스톤 디자인 **인원** : 4명

기간 : 22.07.01 – 23.12.16

기술 :
Rockchip NPU, SRGAN, VNC, RS232, k8s, Jenkins, C/C++

업무

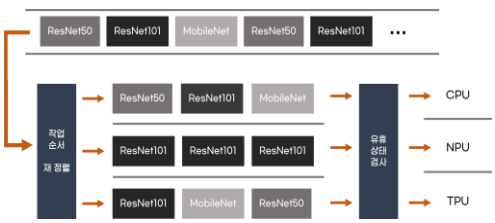
- 인프라에 접속하는 단말기의 네트워크 통신과 연산 비용 관련 메트릭 수집 모듈 개발
- VNC 클라이언트 코드를 Fork하여 Rockhip NPU기반에서 SRGAN이 수행하는 미들웨어 개발

성과

- i7-7700K 서버에서 128명이 동시에 웹 브라우징하는 환경에서도 안정적임
- NPU기반 VNC 클라이언트를 통하여 평균 75 %의 통신 비용 감축함 (8mbps -> 2mbps)
- 기존의 VNC는 2~3 FPS이였지만 통신 비용을 낮춰 통신한 결과 7~8 FPS 로 향상됨

[1] 2023 인공지능 및 응용 워크숍 at 정보통신학회 (우수 논문)
"초해상도 및 기반의 효율적인 가상 데스크톱 인프라 설계"
[2] 클라우드 노트북 (특허 출원 10-2022-0125883)
[3] 소스코드 : 핵심 모듈 코드, 인프라 코드

프로젝트 : 이기종 컴퓨팅과 복수개의 신경망 모델 추론 환경에서 높은 처리량을 위한 스케줄러 관한 연구



연산장치 간 작업 분배하는 스케줄링 알고리즘

내용

이기종 컴퓨팅 환경에서 효율적인 멀티 모달 추론을 위해 연산 장치의 특성 분석과 실시간 작업 분배 스케줄을 통한 성능 분석

기관 : ETRI

기간 : 22.12.20 – 23.02.14

기술 :
TFLite, OpenVINO, Rockhip NPU, OpenCL, C/C++

업무

- 5종류의 연산 장치를 일괄 관리와 통합된 추론환경 개발[Arm CPU][Mali GPU][Intel NCS2][Coral TPU][Rockchip NPU]
- 실시간으로 최적 추론 처리량을 탐색하는 스케줄러 구조 설계

성과

- 연산량이 적은 신경망 모델을 파티셔닝하여 이기종 컴퓨팅을 수행하는 것은 비효율적인 문제 발견
- 연산 장치마다 수행 가능한 데이터 형식이 다른 문제로 인해 역양자화와 데이터 이동이 핵심 병목 문제
- 스케줄링 결과는 작업 분배 알고리즘에 따라 최대 2.7배 성능 개선이 이뤄짐

[1] JIEIE, Vol.60(7), 40-49, July. 2023 (**차주형**, 권용인, 이제민)
"이기종 컴퓨팅과 복수 신경망 추론 환경에서 높은 처리량을 위한 스케줄러 관한 연구"
[2] 소스코드 : https://github.com/Piorosen/HeterogeneousPU