서버-엣지 협업 기반 저지연 다중 사람 포즈 추정 시스템

김량수, 김근용, 김재인, 유 학, 윤기하, 김성창
한국전자통신연구원
rskim@etri.re.kr

Real-time multi-person pose estimation based on server-edge distributed computing

Ryangsoo Kim, Geunyong Kim, Jaein Kim, Hark Yoo, Giha Yoon, Sung Chang Kim
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문에서는 AI칩이 장착된 저가형 임베디드 보드를 활용하여 영상 내 사람을 검출한 뒤 각 사람에 대한 포즈추정을 수행하는 하향식(Top-Down) 방식의 다중 사람 포즈 기술을 구현함에 있어서 무선 네트워크로 연결된 원격 서버의 컴퓨팅 자원을 활용해 추정 소요 시간을 단축시키는 서버-엣지 협업 기반 저지연 다중 사람 포즈 추정 시스템을 소개한다. 임베디드 보드에서는 AI칩을 활용해 Yolov7 객체검출 모델을 추론하여 영상 내 사람으로 인식된 객체의 바운딩 박스 정보를 추출하고 영상 데이터와 바운딩 박스 정보를 무선랜으로 연결된 서버에 전송한다. 서버에서는 수신된 영상 데이터와 바운딩 박스 정보를 활용해 HRNet 포즈추정 모델 추론 연산을 수행하여 검출된 사람별 포즈를 추정한 뒤 결과데이터를 임베디드 보드에 전송한다. 이를 통해 임베디드 보드에 이종 AI 가속기를 탑재하여 다중 사람 포즈를 추정한 시스템보다 본 논문에서 제안하는 서버-엣지 협업 기반 다중 사람 포즈 추정 시스템이 더 빠르게 다중 사람 포즈를 추정할 수 있음을 실험을 통해 확인하였다.

Ⅰ. 서 론

본 논문에서는 라즈베리파이 및 NVIDIA JETSON 보드와 같이 한정적인 컴퓨팅(CPU/GPU) 자원을 보유하고 있는 임베디드 보드에 장착된 카메라에서 수집된 영상 데이터로부터 검출된 사람의 포즈를 실시간으로 추정하는 방법을 소개한다. 본 연구실에서 수행한 기존의 연구에서는 단일 임베디드 보드에 이종 AI 가속기(GPU+AI칩)를 탑재하여 실시간 객체 감지와 다중 사람 포즈 추정을 구현하는 기술이 활용되었다 [1]. 이종 AI 가속기를 사용한 다중 포즈 추정 시스템의 경우, AI칩에서는 Yolov7을 활용한 객체 검출을 수행하고 GPU에서는 검출된 사람 객체에 대하여 HRNet 모델을 활용해 포즈를 추정하는 방식으로 구현되었다. 위 방식의 경우 사람의 수가 많아질수록 GPU를 사용하는 HRNet 모델 추론 소요시간이 비례하여 증가하므로 임베디드 보드의 한정적인 GPU 자원으로는 많은 사람이 영상에 존재 할 경우 다중 사람에 대한 포즈 추정 소요 시간이 증가하는 원인이 된다. 이를 해결하기 위해 본 논문에서는 임베디드 보드와 가까운 위치에 배치 되어있는 서버의 컴퓨팅 자원을 활용해 포즈 추정 모델 추론 연산을 오프로딩하는 서버-엣지 협업 기반 다중 사람 포즈 추정 시스템을 소개한다.

Ⅱ. 엣지-서버 협업 다중 포즈 추정 기술 구성

본 논문에서는 다중 사람 포즈 추정 방법으로 영상 내 사람별로 포즈 추정을 수행하는 하향식(Top-Down) 방식을 고려한다. 하향식 다중 사람 포즈 추정 방식은 객체 검출 딥러닝 모델을 사용해 영상 내 사람 객체의 위치정보(중심좌표 + 높이/넓이)가 포함 되어있는 바운딩 박스(Bounding box) 정보를 추출하고, 해당 바운딩 박스 영역에 포함되는 영상 데이터를

그림 1. 서버-엣지 협업 기반 다중 사람 포즈 추정 시스템 구성도

스크랩(Crop)하여 포즈 추정 딥러닝 모델의 입력 데이터로 사용하여 스크랩 된 영상 데이터 내 단일 사람의 포즈를 추정하는 방식으로 동작한다. 본 논문에서 제안하는 서버-엣지 협업 기반 다중 사람 포즈 추정 시스템은 그림 1에서와 같이 임베디드 보드에서 객체 검출을 수행하는 엣지 AI 기술과 서버에서 사람 포즈를 추정하는 서버 AI 기술의 협력을 통해 구현된다.

엣지AI 기술로는 Hailo사의 Hailo-8 AI칩이 탑재된 NVIDIA Jetson Xavier Nx 임베디드 보드에서 Hailo Datafolow Compiler를 활용해 INT8 연산으로 양자화된 Yolov7 모델(입력사이즈: 640x640)을 HailoRT를 활용해 런타임 엔진을 생성하여 python 프로그램으로 구현하였다. 엣지 AI가 구동되는 임베디드 보드에서는 입력 영상 데이터에 엣지 AI를 적용해 객체 검출을 수행하고, 검출 결과 중 사람 클래스에 해당하는 객체가 검출된 경우 해당 영상 데이터(스틸샷 이미지)와 json 포맷으로 변경된 사람 객체 바운딩 박스 정보를 서버에 전송하여 포즈 추정 연산을 요청한다. 서버 AI 기술로는 NVIDIA GPU(RTX 3070)가 장착된 서버에 PyTorch로 학습 완료된 HRNet 모델(입력데이터: 265x192)을 NVIDIA의 TensorRT 프로그램을 사용해 런타임을 생성하여 python 프로그램으로 구현하였다[2]. 서버 AI를 통해 도출된 사람별 포즈추정 결과 데이터는 json 포맷으로 변환되어 다시 임베디드 서버로 전송되고, 임베디드 보드에서는 서버에서 수신한 json 데이터를 변환하여 기존 영상 데이터 내 각 사람의 포즈 추정 결과 데이터를 이미지에 표시하여 모니터 화면에 출력한다.

위의 서버-엣지 협업에 있어서 서버-클라이언트 사이의 원활한 원격함수 호출을 사용하기 위해서 오픈 소스 RPC 프레임워크인 gRPC를 사용하여 구현하였다. 소프트웨어 통신 프로토콜인 gRPC는 원격 시스템에서 실행되며 로컬 시스템 호출과 매우 유사한 원격 서브루틴의 호출을 용이하게하여, gRPC API를 통해 임베디드 장치가 서버에서 실행되는 서버 AI 활용 사람 포즈 추정 모델 추론 함수를 호출할 수 있도록 합니다. 위 과정에서, 신속한 메시지 직렬화를 위한 이진 메시지 형식인 Protobuf를 사용하면 gRPC 메시지가 직렬화되어 API 호출 중 서버-엣지간 주고받는 데이터의 사이즈가 줄어들어 서버-엣지 협업에서 필연적으로 발생하는 네트워크 지연이 크게 줄어들게 되는 장점이 있다. 위 기술은 python의 RPC 라이브러리를 사용하여 구현되었다.

	엣지 AI (GPU)	엣지AI (AI칩+GPU)	서버-엣지 협업
FPS	1.0	8.7	25.2
Latency(ms)	990.1	758.5	170.5

표 1. 컴퓨팅 환경 변화에 따른 다중 사람 포즈 추정 성능 변화

그림 2. 다중 사람 포즈 추정 실험 결과

Ⅲ. 실험결과

표 1과 그림 2는 본 논문에서 제안하는 서버-엣지 협업 기반 다중 포즈 추정 기술의 성능을 검증하기 위해 수행한 실험 결과이다. 실험에 사용한 영상 데이터는 5명의 사람이 등장하여 안무를 연습하는 MP4 파일을 사용했다. 위 표의 결과에서 알 수 있듯이, 본 논문에서 제안하는 서버-엣지 협업을 통해 다중 포즈 추론 시스템을 구현하면 단일 영상에 대하여 프레임처리성능(FPS, Frame-per-second)이 GPU 가속장치만 사용하는 엣지 AI 대비 약 25배 향상되며 AI칩과 GPU 동시 사용하는 이종 가속장치 활용 엣지 AI 대비 약 3배 향상되었음을 확인하였다. 또한, 단일 프레임에 대하여 다중 포즈 추정 완료까지 소요되는 지연시간(Latency)도 단일 및 이종 가속장치 활용 엣지 AI 대비 약 83%와 77% 단축되었음을 확인하였다. 이를 통해 서버-엣지 사이의 네트워크 연결이 안정적인 환경하에서는 엣지 AI만을 사용하기보다는 서버-엣지 협업을 통해 엣지 AI의 부족한 컴퓨팅 자원을 보충하여 AI 기반 영상분석을 수행하는 방법이 분석 지연시간과 같은 처리 성능을 향상시킴을 확인하였다.

Ⅳ. 결론

본 논문에서는 프레임 처리량과 처리 지연시간을 개선하는 방법으로 하향식 다중 포즈 추정 기술의 구현에 있어서 서버-엣지 협업 기술을 도입하였다. 실험 결과를 통해 입증한 바와 같이, 서버와 엣지 간의 네트워크 연결이 안정적인 환경에서는 제안된 기술을 사용하면 단일 영상에 대한 복잡한 AI 기반 영상분석의 저지연 처리가 가능하다고 예상된다. 향후 연구에서는, 모바일 사용자를 대상으로 복합 AI 모델 기반의 영상분석 서비스를 제공하는 방법으로, 5G 네트워크 환경에서의 서버-엣지 협업 기반 AI 추론 시스템을 ETRI 호남권연구본부의 5G 특화망을 활용하여 구축할 계획이다. 이를 통해 다양한 시나리오에서의 성능 평가를 통해 서버-엣지 협업 기술을 고도화할 예정이다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(RS-2024-00332050, AI기반 개방형 5G-A 융합서비스 테스트베드 구축운영)

참 고 문 헌

[1] 김량수, 김재인, 유학, 김성창, “이종 AI 가속기를 탑재한 임베디드 보드에서의 실시간 객체 감지와 다중 사람 포즈 추정 구현 및 성능 평가,” 2023년도 한국통신학회 하계종합학술발표회, pp.1028-1029, 2023.

[2] Simple-HRNet github, [Online] Available : https://github.com/stefanopini/simple-HRNet