Cython 기반 정수 양자화 기법을 활용한 지상 차량용 경량 비전 시스템의 실시간 최적화에 관한 연구

노범석, 박관익, 백승호

LIG넥스원

beomseok.noh@lignex1.co.kr, gwanik.park@lignex1.com, seungho.baek@lignex1.com

A study on real-time optimization of lightweight vision system for ground vehicles using cython-based integer quantization technique

Beomseok Noh, Wooyeol Hyun, Seungho Baek

요 약

본 논문은 사전 학습된 합성곱 신경망(CNN) 모델을 기반으로, 지상 차량용 임베디드 비전 시스템에서의 실시간 처리를 위한 경량화 시스템 구조를 제안한다. 특히, 부동소수점 연산에 비해 연산 및 메모리 효율이 높은 정수 기반 추론을 목표로, Cython을 활용한 정수 양자화 기반의 연산 최적화 방법을 설계하였다. 모델 학습 이후 적용 가능한 Post-Training Quantization(PTQ) 기법을 바탕으로, CNN의 Conv2D 및 FC 계층을 8비트 정수 연산으로 변환하고, 이에 적합한 Cython 구현체를 통해 CPU 환경에서도 실시간 처리 성능을 확보할 수 있는 시스템 구조를 제안한다.

Ⅰ. 서 론

기존의 딥러닝 추론 시스템은 대부분 GPU 나 NPU 등의 고성능 하드웨어를 전제로 개발되어, 실제 저전력 임베디드 시스템이나 군용 차량, 무인 전술 차량 등의 제한된 환경에 바로 적용하기 어렵다. 특히, 국방 및 방산 분야에서는 신뢰성과 보안, 전력 효율성을 고려한 경량화된 AI 시스템에 대한 요구가 지속적으로 증가하고 있다. 양자화(Quantization)는 이러한 문제를 해결할 수 있는 대표적인 기법으로 부동소수점 기반의 합성곱 신경망을 정수 기반으로 변환함으로써 모델 크기와 연산량을 줄이고, CPU 만으로도 실시간 처리가 가능하게 경량화 할 수 있다.[1][2]

그러나 기존의 정수 양자화 기법은 TensorRT, TFLite 등의 프레임워크에 제한되어 있으며, 이들 대부분은 특정 하드웨어나 OS에 종속적이고, 필요한 상황에 맞게 수정이 어렵다는 단점이 존재한다.[3] 이에 본 연구는 고정 하드웨어에 종속되지 않으면서도 실시간 처리가 가능한 양자화된 추론 시스템을 Python 기반으로 구현하는 것을 제안하고자 한다.

특히, Python과 C의 장점을 결합한 Cython 구현체를 이용하여 합성곱 신경망의 주요 연산을 정수화된 형태로 직접 구현하고, 이를 통해 GPU 없이도 지상 차량에서 실시간 처리가 가능한 경량 비전 시스템 구조를 제안한다.[4]

Ⅱ. 본론

본 논문에서 제안하는 시스템은 지상 차량용 경량 비전 시스템을 목적으로 하며, 사전 학습된 합성곱 신경망 모델을 기반으로 최적화 하여 제한된 연산 자원에서도 실시간 처리가 가능하도록 입력 처리 계층, 추론 계층, 출력 및 후처리 계층과 같이 3계층 아키텍처로 구성된다.

2.1 지상 차량용 경량 비전 시스템

2.1.1 입력 처리 계층(Input Preprocessing)

입력 처리 계층은 차량에 장착된 카메라나 센서로 부터 입력되는 실시간 영상 데이터를 수집하고, CNN 추론에 적합한 형태로 전처리한다. 주요 기능은 영상 리사이징, 픽셀 정규화 및 8-bit 정수로의 클리핑 ,채널 정렬 ,배치 입력 구성을 위한 프레임 버퍼링이다.

이 계층의 전처리 파이프라인은 사용자의 온보드 시스템에 적합한 언어를 사용하여 구현 가능하며 Python을 사용할 경우 Numpy 또는 OpenCV로 구현 할수 있고, 이후 CNN 입력으로 적합한 np.ndarray[np.int8_t] 형태로 변환할 수 있다.

2.1.2 추론 계층 (Quantized Inference)

추론 계층은 본 시스템의 핵심으로, 정수 양자화된 CNN 연산을 Cython으로 직접 구현하여 CPU 상에서도 빠른 추론이 가능하도록 최적화한다. 추론 계층은 Conv2D연산, ReLU 및 Activation연산, FC(Fully Connected) Layer연산을 Cython 구현체를 활용하여 연산한다.

Conv2D 연산 과정에서는 cython.parallel 또는 loop unrolling 최적화가 적용된 im2col-free 방식의 int8 기반 합성곱 연산을 수행하고 ReLU 및 Activation 연산에서 상한 클리핑을 통해 8-bit 정수로 변환을 수행한다. Fully Connected Layer 연산에서는 Cython 변수를 선언하여 활용함으로써 C-Level 수준의 연산성능을 확보한다.

이렇게 작성된 추론 모듈은 Python 코드에서 직접 호출 가능하며, cythonized .so 형태로 빌드되어 RTOS나 임베디드 환경에 손쉽게 통합할 수 있다.

2.1.3 출력 및 후처리 계층 (Postprocessing & Visualization)

출력 및 후처리 계층은 추론 결과로 생성된 클래스 확률, 바운딩 박스, 위치 좌표 등의 정보를 후처리하여 사용자에게 직관적으로 시각화한다. 이 계층의 주요 역할은 softmax, argmax 등의 후처리 연산, 바운딩 박스

재정렬 및 비정상 값 제거, 결과를 이미지 위에 시각화하여 GUI로 송출이다.

이 계층은 사용자의 온보드 시스템에 적합한 언어를 사용하여 구현이 가능하며 리눅스 기반 GUI 또는 차량 내 디스플레이에 연동 가능하며, 전후방 탐지 시스템, 자율 주행 판단 모듈 등과도 인터페이스할 수 있다.

2.2 기대 효과

2.2.1 CPU 환경에서 실시간 처리성능 확보

사전 학습된 합성곱 신경망(CNN) 모델을 기반으로 CPU환경에서 동작하는 지상차량용 임베디드 비전 시스템을 구현하기 위해서 Post-Training Quantization(PTQ)기법을 사용하여 학습된 모델의 가중치, 추론의 부동소수점(FP32) 연산에 비해 연산 및 메모리 효율이 높은 정수(INT8) 추론 연산을 통해 성능을 확보함을 전제로 한다.

그림1은 본 논문에서 제안한 시스템 구조에 대한 흐름도에 대한 내용이며 PTQ기법을 활용하여 양자화 된 가중치와 모델의 구조를 동일하게 사용하며 2.3에서 기술한 추론 계층에서 사용되는 연산의 과정을 Cython 구현체를 활용하여 실시간 성능 확보를 제안한다.

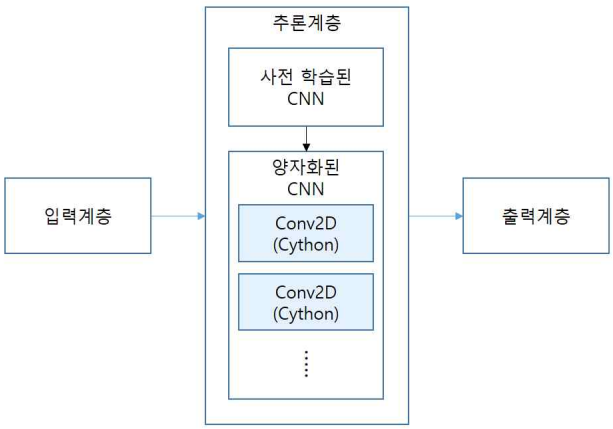

그림 1 제안 방법 흐름도

제안하는 시스템 구조는 기존 프레임워크(TensorFlow/PyTorch)에서 Numpy기반으로 구현되어 있는 Conv2D 연산대비 표1의 기대효과를 갖는다.

항목	기존 프레임워크 (FP32)	Cython 기반 INT8 최적화
연산 정확도	높음 (float 기반)	손해 있음 (정밀도 감소)
연산 속도 (CPU)	느림 (CPU-only)	빠름 (int8 x int8 → int32 연산 최적화)
메모리 사용량	FP32	INT8 (FP32기준 1/4)
에너지 효율	낮음	높음 (CPU 환경에서)

표 1 Cython 기반 최적화 기대효과

이러한 기대효과를 검증하기 위하여 시스템의 핵심 구조인 Conv2D 연산을 Cython 구현체를 활용하여 최적화 하여 벤치마크 실험을 진행하다. 벤치마크 실험에서는 128*128의 입력 값을 출력 채널수 16으로 하여 1번 연산한 값을 CPU-only환경에서 실험하였고, 이때 커널은 3*3 채널 값은 3으로 설정하여 실험하였다. 해당 벤치마크 실험에 대한 결과는 표2와 같다.

항목	기존 프레임워크 (Numpy)	Cython 기반 INT8 최적화
FP32 연산	6.7291 초(s)	2.6676 초(s)
INT8 연산	6.4982 초(s)	1.5091 초(s)

표 2 Cython 기반 최적화 벤치마크 실험

Ⅲ. 결론

본 논문에서는 GPU 나 NPU 같은 고성능 하드웨어가 없는 환경에서도 실시간 비전 처리가 가능한 Cython 기반 정수 양자화 경량 비전 시스템 구조를 제안하였다. 제안된 시스템은 Python 기반의 생산성과 C 언어 수준의 성능을 모두 활용할 수 있으며 양자화 기법 활용 시 특정 하드웨어나 OS에 대한 종속성을 해결 할 수 있고 cythonized .so 형태로 빌드 되어 RTOS 또는 임베디드 환경에서도 손쉽게 통합이 가능하다. 또한 연산 처리속도 향상을 통해 CPU-only 환경인 군용 지상 차량이나 전력 제한형 임베디드 시스템에서도 안정적인 동작이 가능 할 것으로 기대된다.

이러한 시스템 구조를 통해 군용 지상 차량이나 전력 제한형 임베디드 시스템에서의 한계를 극복할 수 있다면, 경량 비전 인식 시스템의 민간/군복합 응용 확대 및 로우 파워 엣지 AI 기술 확산에 기여할 수 있을 것으로 기대된다.

참 고 문 헌

[1] Han, Song, et al. "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding." International Conference on Learning Representations (ICLR), 2016.

[2] Jacob, Benoit, et al. "Quantization and training of neural networks for efficient integer-arithmetic-only inference." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[3] Ahn, Hyunho, et al. "Performance characterization of using quantization for dnn inference on edge devices: Extended version." arXiv preprint arXiv:2303.05016 (2023).

[4] A. Milla and E. Rucci, "Performance Comparison of Python Translators for a Multi-threaded CPU-bound Application," arXiv preprint arXiv:2203.08263 (2022).

[5] Jung, Sangil, et al. "Learning to quantize deep networks by optimizing quantization intervals with task loss." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[6] Mundichipparakkal, Junaid, et al. "Evaluation of Quantization Methods for Neural Networks." (2022).

[7] Gonzalez, R. C., et al. "Albawi, S., Mohammed, TA, & Al-Zawi, S.(2017). Understanding of a Convolutional Neural Network. International Conference on Engineering and Technology (ICET). Antalya, Turkey: IEEE. Blake, JH, Keinath, AP, & Kluepfel, M.(2018, Desember 13). Tomato Diseases & Disorders. Retrieved from Clemson Cooperative Extension: Home &." Data Engineering 122: 127.