딥러닝 모델 성능 최적화를 통한 단일 보드 컴퓨터 기반 실시간 동공 추적 시스템 구현

최건호, 김석찬*

부산대학교, *부산대학교

cgh2022@pusan.ac.kr, *sckim@pusan.ac.kr

Implementation of Real-Time Pupil Tracking System based on Single Board Computer through Deep Learning Model Performance Optimization

Geonho Choi, Suk Chan Kim*

Pusan Nat'l Univ., *Pusan Nat'l Univ.

요 약

동공 추적 기술은 아이트래커의 시선 추적이나 운전자 졸음 방지 등의 응용에 활용되는 핵심 기술로 최근 활발히 연구되고 있다. 동공 추적 기술을 활용하는 시스템은 딥러닝 모델의 빠른 추론 속도와 높은 정확도를 요구한다. 이에 따라 대부분의 연구는 고성능 모델 개발에 초점을 맞추고 있다. 하지만 고성능 모델을 동작시키기 위해선 고성능 데스크탑이 요구되는데 이는 비용, 전력 소모, 이동성 등이 고려되어야 하는 실제 응용 환경에 적합하지 않다. 따라서 동공 추적 시스템 구현을 위해서는 실제 응용 환경에 적합한 단일 보드 컴퓨터와 딥러닝 모델 최적화 과정이 필요하다. 본 연구에서는 딥러닝 모델의 성능 최적화를 통해 단일 보드 컴퓨터 기반 실시간 동공 추적 시스템을 구현한다. 성능 최적화된 모델은 추론 속도 최적화 및 혼합 정밀도 기법을 사용하여 동공 추적 모델의 추론 속도를 향상시키고, 후처리 알고리즘의 파라미터 조정을 통해 정확도를 향상시킨다. 성능 최적화된 동공 추적 모델은 단일 보드 컴퓨터에서도 데스크탑에서 동작시켰을 때와 유사한 정확도와 실시간으로 동공 추적이 가능한 추론 속도를 가진다.

Ⅰ. 서 론

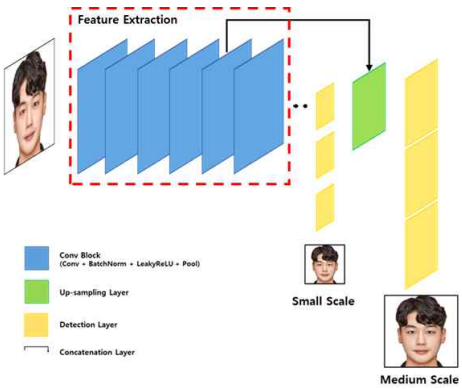
동공 추적 기술은 아이트래커의 시선 추적이나 운전자 졸음 방지 등의 응용에 활용되는 핵심 기술 중 하나로, 최근 활발히 연구되고 있다[1]. 일반적으로 동공 추적 기술을 활용하는 시스템의 경우, 딥러닝 모델의 빠른 추론 속도와 높은 정확도를 요구한다. 대부분의 연구는 고성능 모델 개발에만 초점을 맞춰왔다. 고성능 모델은 추론 과정에서 많은 연산량을 처리하기 위해 고수준 그래픽 처리장치(GPU)가 장착된 데스크탑이 요구되는데, 데스크탑은 높은 비용 및 전력 소모, 낮은 이동성 등의 제약을 가지기 때문에 실제 응용 환경에서 사용하기 적합하지 않다. 따라서 실제 응용 환경에서 실시간 동공 추적 시스템 구현을 위해선 저전력으로 구동되는 소형 단일 보드 컴퓨터와 이에 맞는 딥러닝 모델 성능 최적화가 필요하다.

단일 보드 컴퓨터는 데스크탑이나 개인용 컴퓨터와 달리 단순한 아키텍처를 가지는 시스템으로, 자율/이동 시스템을 목표로 하는 딥러닝 응용 분야에 적합하다. 하지만 연산량이 많은 고성능 모델을 직접적으로 온보딩 했을 때에는 성능을 보장할 수 없다. 이 문제를 해결하기 위해선 온보딩 할 모델의 최적화 과정이 필수적이다.

본 논문에서는 딥러닝 모델 성능 최적화를 통해 단일 보드 컴퓨터 기반 실시간 동공 추적 시스템을 구현한다. 단일 보드 컴퓨터로는 NVIDIA 젯슨 나노를 사용하였으며, 전이 학습을 통해 동공 추적 모델을 생성했다. 최적화 기법에 따른 동공 추적 모델의 성능을 평가하였으며, 성능 최적화를 통해 단일 보드 컴퓨터에서도 실시간 동공 추적 시스템 구현이 가능함을 확인했다.

Ⅱ. NVIDIA Jetson Nano

NVIDIA 젯슨 나노는 NVIDIA 젯슨 나노는 저전력 소비를 특징으로 하는 단일 보드 컴퓨터로 임베디드 어플리케이션, 딥러닝, 사물 인터넷, 컴퓨터 비전 등을 위해 개발된 JetPack SDK와 딥러닝 추론 최적화를 위한 Tensor Real Time(TensorRT) 라이브러리를 포함하고 있다[2]. 젯슨 나노는 이미지 분류, 객체 탐지, 세분화 및 음성 처리 등의 딥러닝 응용에 주로 사용된다.



〔그림 1〕 YOLOv3-tiny의 구조

Ⅲ. 객체 탐지 모델

딥러닝 기반 동공 탐지 모델로 YOLOv3-tiny를 사용했다[3]. YOLOv3-tiny는 객체 위치 추정(Localization)과 분류(Classification), 두 과정을 동시에 수행하는 1-stage detector로 높은 추론 속도를 가진다. 모델은 합성곱 신경망을 통해 주요 특징을 추출하고 FPN(Feature Pyramid Network)구조를 가진 분류기의 다중 스케일 특성맵을 활용하여 객체를 인식한다. [그림 1]은 YOLOv3-tiny 모델의 구조를 나타낸다.

Ⅳ. 모델 훈련 및 최적화 과정

GeForce RTX 2080Ti가 GeForce RTX 2080Ti가 장착된 고성능 데스크탑에서 전이 학습을 통해 기초 모델을 생성했다. 학습된 모델을 젯슨 나노에 임베딩 하기 위해 TensorRT 기반 딥러닝 추론 최적화를 진행했다.

모델 훈련 및 성능 평가를 위한 데이터를 수집하기 위해 웹 상에서 이미지를 가져와서 저장하는 이미지 크롤링(Image Crawling) 기법을 사용했다. 수집한 데이터는 눈과 동공이 잘 보이는 정면 응시 사진이다. 모델의 정확도를 향상시키기 위해 데이터 증강기법을 사용했다. 또한, K-평균 클러스터링(K-means clustering) 기법을 적용하여 이미지 내의 객체들의 크기를 가장 잘 대변하는 앵커 박스를 선정했다.

젯슨 나노에서 딥러닝 모델을 동작시키기 위해선 최적화 과정이 필수적이다. TensorRT를 활용하여 학습된 딥러닝 모델의 추론 최적화를 진행했다. 추론 최적화 과정에는 양자화 및 정밀도 교정, 그래프 최적화, 커널 자동 조정, 동적 텐서 메모리 및 다중 스트림 실행 등이 있다. 추론 과정에서 더 빠르고 효율적인 연산을 위해 32비트 부동 소수점 형식을 사용하는 대신 정밀도가 낮은 16비트 부동 소수점 형식을 혼합하여 사용했다.

NMS 알고리즘은 모델이 예측한 경계 상자 중 가장 적합한 상자를 선택하는 후처리 과정으로, SSD, YOLO를 포함한 대부분의 객체 탐지 모델에서 활용된다[4]. NMS 알고리즘에서 IoU(Intersection of Union) 임계값 설정에 따라 탐지 모델의 정확도 성능은 상이할 수 있다. 따라서 탐지하고자 하는 객체의 특징과 탐지 모델의 추론 방식에 따라 적절한 IoU 임계값 설정이 필요하다. 본 연구에서는 모의 실험을 통해 동공 탐지에 적합한 IoU 임계값을 도출하였다. 해당 결과를 모의 실험 결과에서 분석한다.

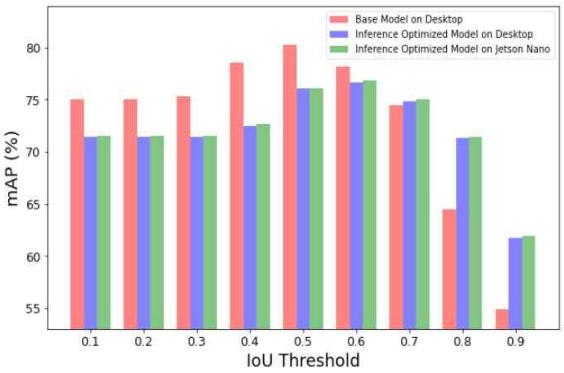
Ⅴ. 모의 실험 결과

본 장에서는 이미지 크롤링으로 수집한 196개의 이미지와 Logitech BRIO 웹캠으로 촬영한 실시간 영상을 사용하여 동공 탐지 모델의 성능을 분석했다. 모델의 정확도와 추론 속도의 평가지표로 mAP(mean Average Precision)와 FPS(Frame Per Second)를 사용했다.

IoU 임계값에 따른 mAP 성능을 그림 4에서 비교했다. 최적화 과정을 진행하지 않은 기초 모델의 경우, IoU 임계값이 0.5일 때 최대 mAP 성능을 보였다. 반면, 추론 최적화가 진행된 모델에서는 IoU 임계값이 0.6일 때 최대 성능을 보였다. 따라서, 추론 속도 최적화를 적용한 모델에서는 0.6의 IoU 임계값을 적용하는 것이 적합하다..

[표 1] 최적화 기법 및 장치에 따른 모델 성능

Hardware	TensorRT Optimization	Mixed Precision	mAP (%)	FPS
Desktop with RTX 2080 Ti	x	x	78.18	59
	o	x	76.62	228
	o	o	76.62	277
Jetson Nano	x	x	ERR	ERR
	o	x	76.61	17
	o	o	76.81	25



〔그림 4〕 NMS 알고리즘 적용 시 IoU 임계값에 따른 mAP 비교

GeForce RTX 2080Ti가 장착된 데스크탑과 젯슨 나노에서 최적의 IoU 임계값이 적용된 딥러닝 모델을 동작시켰을 때, 최적화 기법에 따른 동공 탐지 성능을 표 2에 나타냈다. 최적화 과정을 진행하지 않은 기초 모델은 80.26% mAP와 59 FPS 성능을 보였다. 기초 모델에 TensorRT 최적화만 수행했을 때, mAP는 약 4.5% 하락하였지만 FPS는 약 3.86배 상승했다. 혼합 정밀도(Mixed Precision) 기법을 추가적으로 적용했을 때는, mAP의 변화 없이 약 1.22배의 FPS 성능 개선을 보였다. 젯슨 나노에서 추론 최적화를 수행하지 않은 모델을 동작시킨 경우, 메모리 에러로 인해 성능 측정이 불가능했다. 그러나 TensorRT 최적화를 진행한 모델은 젯슨 나노에서 동작 가능했으며, 76.61% mAP와 17 FPS 성능을 보였다. 혼합 정밀도 기법을 추가적으로 적용했을 때는, 비슷한 수준의 mAP와 약 1.47배 향상된 FPS 성능을 보였다. 모의 실험에서 도출한 최적의 IoU 임계값과 추론 속도 최적화가 적용된 동공 추적 모델은 단일보드 컴퓨터에서 동작되었음에도 불구하고, mAP 성능에서 적은 하락을 보였으며, 실시간으로 동공 추적이 가능한 25 FPS 성능을 보였다.

Ⅵ. 결론

본 논문에서는 단일 보드 컴퓨터에서 실시간 동공 추적 시스템 구현을 위해 딥러닝 모델의 정확도 및 추론 속도를 최적화했다. 최대 정확도 성능을 내는 최적 IoU 임계값을 도출했으며, 추론 속도 최적화를 위해 TensorRT 최적화 및 혼합 정밀도 기법을 사용했다. 최적화 기법에 따른 mAP와 FPS 성능 비교를 통해, 단일 보드 컴퓨터에서 실시간 동공 추적 시스템 구현이 가능함을 확인했다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2022R1A2C1092737)

참 고 문 헌

[1] W. Fuhl, et.al. “Pupilnet: Convolutional neural networks for robust pupil detection,” *In CoRR*, 2016.

[2] Il-Sik Chang, Gooman Park, “Implementation of Jetson Nano Based Face Recognition System,” JKICS, 2021.

[3] Joseph Redmon, Ali Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[4] A. Neubeck and L. Van Gool, "Efficient Non-Maximum Suppression," *18th International Conference on Pattern Recognition (ICPR'06)*, 2006.