

임베디드 AI 장치에서 차선과 차량을 동시에 검출하는 경량 멀티 태스크 CNN

김태현, 박민서, 신현식, 김형원*
*충북대학교

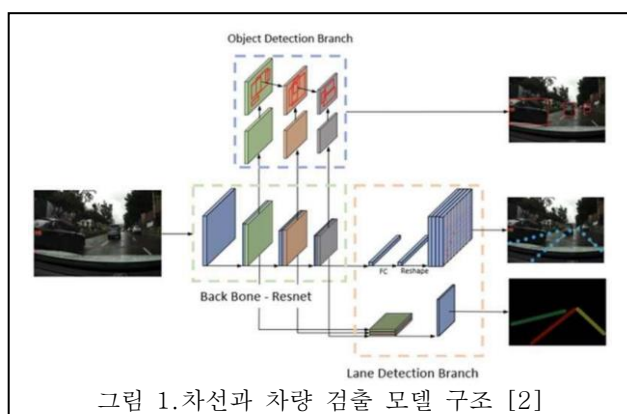
thk4627@chungbuk.ac.kr, minseo@chungbuk.ac.kr, hsix@chungbuk.ac.kr, *hwkim@cbnu.ac.kr

Light-Weight Multi-Task CNN for Concurrent Detection of Lanes and Vehicles On Embedded AI Devices

Kim Tae Hyeon, Park Min Seo, Shin Hyeon Sik, Kim Hyung Won*
Chungbuk National University

요 약

첨단 운전자 보조 시스템(Advanced Driver Assistance Systems)이 지속적인 발전을 함으로서 차선 검출과 차량 검출 기술의 중요성 또한 증가 하였다. 최근에 차선 및 차량 검출의 높은 정확도를 달성하기 위해 다양한 CNN 모델들이 개발되었으며, 개별 CNN 모델들의 정확도와 속도 향상을 위한 많은 연구가 추진되었다. 그러나 차선 검출과 차량 검출을 수행하는 각각의 CNN 모델들을 구현한 기존 기술은 파라미터 수가 크게 증가하여 ADAS 를 위한 소형 임베디드 AI 장치에 탑재하기 어려운 문제를 가지고 있으며, 또한 추론 속도가 낮은 제약 사항을 가지고 있다. 이를 해결하기 위해 본 논문에서는 차선 검출 모델과 차량 검출 모델의 Feature extraction 백본을 공유[1][2]하고 서로 상이한 layer 들을 Multi-Head 로 구현하는 멀티 태스크 CNN 모델을 구현하고 성능을 분석한다. 구현된 멀티 태스크 CNN 모델의 네트워크 크기를 더욱 줄이기 위해 TensorRT[3] tool 을 활용하여 네트워크 최적화 및 파라미터 Quantization 을 수행한다. 최적화된 멀티 태스크 CNN 모델을 상용 임베디드 AI 가속기 모듈에 탑재하여 속도와 정확도 분석 결과를 보인다.



I. 서 론

현재 AI 에 관한 사람들의 관심이 급격히 증가하고 있다. 이에 따른 다양한 인공지능 분야에 관한 연구 또한 진행되고 있다. 이중 자율주행 자동차는 미래를 이끌 핵심 기술로 판단되어 연구가 꾸준히 지속되고 있다. 우리는 이 자율주행 자동차에 관해 더 가벼운 모델을 구현하여 적은 파라미터로 동작할 수 있게 구현하여 저전력 장치에서 최적화된 성능을 목표로 연구를

진행하였다. 기존은 차량 검출 그리고 차선 검출에서 많은 파라미터를 사용하여 더 정확한 결과를 목표로 하였다면 우리는 이 두 모델을 하나로 통합하여 파라미터 수를 약 50% 이상 감소시키는데 성공하였고 이를 추가로 최적화 하는 과정을 거쳐 속도를 더욱 개선시켰다.

II. 본론

1) 다중작업 CNN

앞서 언급했듯이 우리의 모델은 백본으로 Resnet[4]을 공유해서 사용한다. 그리고 이 백본에서 YOLO[5] 방식을 사용해서 세단계의 feature 을 뽑아내서 두개의 서로 다른 헤드에 넣어준다. 헤드는 각각 YOLOv3[6]와 UltraFastLaneDetection[7] 헤드로 기존에 있고 속도와 정확도 면에서 강점이 있는 모델을 사용하였다. 이 구조는 그림 1 을 보면 확인할 수 있다.

저전력 장치에선 수용가능한 파라미터 수가 정해져 있다. 따라서 그림 1 의 모델을 기반으로 파라미터 수를 계산 하였고 기존의 YOLOP + UFLD 의 파라미터에 비해 약 50%가 감소한 것을 확인할 수 있었고 이에



따라 저전력 장치에서 감안할 수 있는 연산양으로 구성되었다.

2) 저전력 장치 사용을 위한 TensorRT

우리는 이 적은 수의 파라미터에 더해 추가적인 연산양 감소와 그에 따른 속도 증가를 위한 최적화 방법으로 TensorRT 를 사용하였다. 우리는 기존에 공유하는 백본 부분을 TensorRT 로 변환시켜 계산하였고 뒤의 헤더 부분은 기존의 파라미터를 사용하여 정확도 유지를 목표로 하였다 이에 따라 유의미한 속도 변화가 있었다.

III. 결론

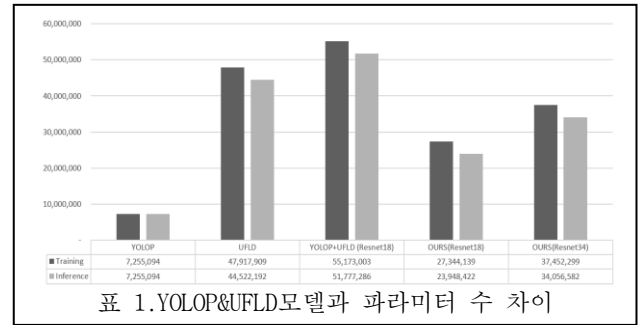
우리는 앞서 말했듯이 파라미터 수를 감소시켰다. 표 1 을 보면 YOLOP 와 UFLD 의 각각의 파라미터 수와 같이 실행 하였을 때의 파라미터 수를 확인할 수 있다. 그리고 우리가 구현한 모델의 파라미터 수 또한 확인할 수 있다. 그 결과로는 약 50%의 파라미터 수가 감소한것을 확인할 수 있었다.

표 2 는 저전력 장치인 Jetson NX 에서 우리의 모델을 실행시킨 결과이다. 시험 데이터로는 CULane[8]을 사용하였고 백본부분에 TensorRT 를 적용시킨 결과 또한 포함시켰다. 기존 모델은 차량은 70 의 mAP, 차선은 0.661 f1-score 를 가지고 1 초당 약 17 의 장수를 동작 시킬 수 있는 것을 확인할 수 있었다. TensorRT 를 적용시킨 결과는 기존 모델과 비교해서 정확도 면에서는 거의 동일한 값을 가졌고 속도 면에선 약 0.1 초 더 빠르게 동작하는 것을 확인할 수 있었다.

위의 모델에 따른 결과는 그림 2 를 통해 볼 수 있듯이 사람이 인식하는 부분과 크게 다르지 않음을 확인할 수 있었고 차량과 차선 모두 정확히 찾는 것을 확인할 수 있었다.

ACKNOWLEDGMENT

This work was supported by Regional Leading Research Center (RLRC) of the National Research Foundation of Korea (NRF) grant funded by the Korean government(MSIT) (No. 2022R1A5A8026986), and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea



Backbone	mAP50(%)	F1-Score	Run time(s)	FPS
Pytorch	70	0.661	0.0581	17.2
TensorRT	70	0.66	0.0492	20.3

표 2. 정확도 및 속도 결과

government(MSIT) (No.2020-0-01304, Development of Self-learnable Mobile Recursive Neural Network Processor Technology). It was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Communication Technology Research Center support program (IITP-2022-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). The correspondence author is Pf.Kim Hyungwon.

참 고 문 헌

- [1] Dong Wu, Manwen Liao, Weitian Zhang, Xinggang Wang, Xiang Bai, Wenqing Cheng, WenyuLiu. YOLOP: You Only Look Once for Panoptic Driving Perception. arXiv: 2108. 11250, 2021.
- [2] Shin Hyeonsik. Multi-Task CNN Model for Concurrent Detection of Vehicles and Lanes. In Preceedings of the Korea Institute of Information and Communication Engineering (KIICE). 2022
- [3] NVIDIA TensorRT. Accessed: Mar. 23, 2021. [Online] Available: <https://developer.nvidia.com/tensorrt>
- [4] KaimingHe,XiangyuZhang,ShaoqingRen,JianSun. Deep Residual Learning for Image Recognition. In Preceedings of the IEEE conference on computer vision and pattern recognition. 2016
- [5] Joseph Redmon,Santosh Divvala,Ross Girshick,Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In Preceedings of the IEEE conference on computer vision and pattern recognition. 2016
- [6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 2
- [7] Zequn Qin, Huanyu Wang, and Xi Li. Ultra fast structure-aware deep lane detection. In Proceedings of the European Conference on Computer Vision (ECCV), pages 276– 291, 2020.
- [8] CULane: a large scale challenging dataset for academic research on traffic lane detection. Introduced by Pan et al. inSpatial As Deep: Spatial CNN for Traffic Scene Understanding. 2017