자율주행 트랜스포머 워크로드를 위한 엣지 AI 가속기와 매핑 전략

국설헌, 조예령, 김중헌
고려대학교
oliviakuk7@gmail.com, joyena0909@korea.ac.kr, joongheon@korea.ac.kr

Edge AI Accelerators for Transformer Workloads in Autonomous Driving: A Survey of System Constraints and Mapping Strategies

Seolheon Kuk, Yeryeong Cho, Joongheon Kim
Korea Univ.

Abstract

본 논문은 자율주행 시스템 내 트랜스포머(Transformer) 기반 워크로드의 추론(inference)을 지원하기 위한 하드웨어 가속기 활용과 시스템적 제약을 다룬다. 최근 자율주행 perception과 planning 영역에서 BEVFormer, trajectory transformer와 같은 모델이 활용되며 높은 성능을 달성하였으나, 연산량, 메모리 대역폭, 전력 제약 문제로 인해 차량 내 임베디드 플랫폼에서의 실시간 처리에 큰 도전이 존재한다. 본 연구는 기존의 트랜스포머 추론 최적화 기법, 에너지 효율을 위한 엣지 AI 연구, 자율주행 소프트웨어·하드웨어 스택 리뷰, 트랜스포머와 하드웨어 가속기의 관계를 기반으로, 자율주행 특화 태스크별 워크로드 - 가속기 매핑(workload - accelerator mapping)을 정리한다. 이를 통해 현 시스템의 병목을 규명하고, 소프트웨어 최적화와 하드웨어 선택을 결합한 설계 방향을 제시하며, 향후 chiplet SoC, in-memory computing, 분산 추론 전략과 같은 미래 연구 과제를 논의한다.

Ⅰ. 서론

트랜스포머(Transformer) 모델은 멀티헤드 어텐션 구조를 통해 장거리 의존성과 멀티센서 융합을 효과적으로 처리하며, BEVFormer, TransFusion 등은 자율주행 perception과 trajectory prediction에서 기존 CNN 모델 대비 우수한 성능을 보여주었다 [4-6]. 그러나 이러한 모델은 대규모 행렬 연산과 메모리 접근을 필요로 하여, 차량 내 엣지 플랫폼에서는 전력(수십~100W), 메모리(수 GB), 지연(<50ms) 제약이 병목으로 작용한다 [1-3]. 기존 연구는 모델 경량화, 양자화, FPGA 기반 가속기 설계 등을 통해 성능 - 전력 균형을 개선하고자 하였으나 [2-5], 대부분 개별 최적화 기법에 집중되어 있어 자율주행 태스크 특성과 하드웨어 자원의 매핑 관계를 포괄적으로 정리한 연구는 부족하다. 따라서 소프트웨어 수준의 개선을 넘어, 하드웨어 아키텍처와 시스템 제약을 통합적으로 고려한 분석이 필요하다. 본 논문은 자율주행 파이프라인 속 트랜스포머 워크로드의 특성과 GPU·FPGA·ASIC/NPU·SoC의 역할을 비교하고, 태스크별 워크로드 - 가속기 매핑 전략을 제시한다. 이를 통해 트랜스포머 기반 자율주행 추론의 병목을 규명하고, 향후 chiplet SoC 및 in-/near-memory computing 설계를 위한 연구 방향을 제안한다.

Ⅱ. 자율주행 시스템

자율주행 시스템은 센서 입력, 소프트웨어 스택, 하드웨어 플랫폼, 출력 액추에이터로 구성된다 [1]. 카메라, LiDAR, 레이더 등은 환경을 인식하고, perception - localization - planning - control 모듈을 거쳐 차량 제어로 이어진다 [2]. 그림 1과 같이, 전체 파이프라인은 센싱부터 제어까지 계층적으로 연결되며, 그림 2와 같이 “Sense - Think - Act” 모델로 단순화할 수 있다 [1-2]. 최근에는 BEVFormer, TransFusion 등 트랜스포머 기반 모델이 이러한 파이프라인 속에서 핵심 역할을 담당한다 [4-6]

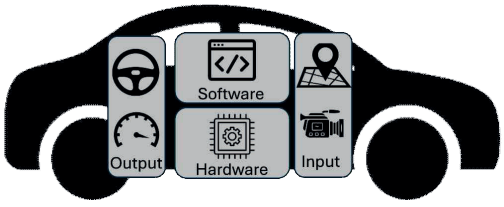

그림 1. 자율주행 시스템 아키텍처

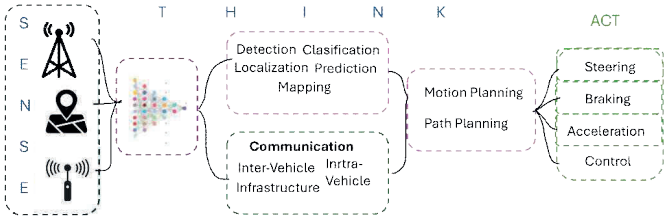

그림 2. Sense-Think-Act 모델

Ⅲ. 트랜스포머 기반 자율주행 하드웨어 시스템

짧은 추론 시간과 낮은 전력 소모는 특히 트랜스포머 기반 워크로드를 처리하는 자율주행 시스템에서 중요한 요구사항이다 [1,6]. 그림 3과 같이 Tesla의 아키텍처는 전처리 모듈(Rectify, RegNet, BiFPN 등) 이후 트랜스포머 블록을 통해 BEV 변환과 특징 융합을 수행하며, 이어 Object Detection, Traffic Light 인식, Lane Prediction 등 다수의 태스크를 병렬로 실행한다 [5]. 이러한 데이터 흐름은 CPU 단독으로는 처리량이 부족해 반드시 가속기의 지원이 필요하다.

GPU는 대규모 행렬 연산과 어텐션 메커니즘을 병렬적으로 처리할 수 있어 BEVFormer, TransFusion 같은 모델 실행에 적합하다. 예를 들어

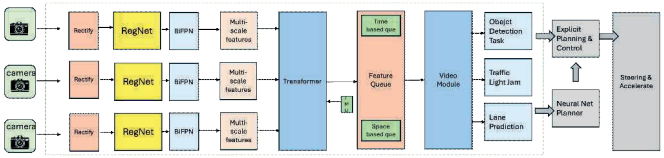

그림 3 테슬라 자율주행 시스템

Apollo 플랫폼은 RTX 3070 GPU를 탑재한 산업용 컴퓨터를 통해 최대 20 TFLOPS 성능을 제공한다 [1]. 그러나 실제 실험에서는 입력 해상도가 높아질수록 CPU에서 발생하는 스케줄링 지연으로 전체 FPS가 저하되는 현상이 관찰되었다 [1]. 또 다른 연구에서는 GPU 기반 추론의 전력 소모가 200W를 초과하면서 차량 ECU 전력 한계를 넘어섰고, 열 관리 문제로 인해 장시간 운행 시 성능 저하(thermal throttling)가 발생하는 것이 확인되었다[4].

FPGA는 데이터 병목현상을 일부 해소할 수 있는 대안이다. Pony.ai의 실험에서는 FPGA를 활용해 CPU를 거치지 않고 GPU로 데이터를 직접 전송했을 때 평균 추론 지연이 약 15% 감소하는 결과가 확인되었다 [1]. 또한 Apollo 프레임워크 기반 연구에서는 Zynq Ultrascale+ FPGA를 이용해 차선 인식과 신호등 탐지를 가속했을 때, GPU 대비 전력 소모가 30% 이상 절감되면서도 유사한 정확도가 유지되는 결과가 나타났다. 그러나 트랜스포머 특유의 동적 연산 패턴을 FPGA에 최적화하는 데는 여전히 한계가 존재한다 [6].

SoC 기반 연구도 활발하다. Tesla의 FSD 칩은 40W 전력으로 72 TOPS 연산을 제공하며, 실제 도로 주행 실험에서 CNN 기반 태스크와 경량 Transformer 블록을 동시에 처리할 수 있음이 검증되었다 [1]. NVIDIA DRIVE Orin은 벤치마크 테스트에서 254 TOPS를 달성했으며, LPDDR5 메모리 인터페이스를 통해 멀티 카메라 트랜스포머 추론을 실시간으로 지원하는 것으로 나타났다 [4]. Mobileye EyeQ 시리즈 또한 heterogeneous 가속기 구조를 도입하여 CNN과 Transformer 연산을 분리 실행했을 때 평균 지연이 20ms 이하로 유지되는 성능이 확인되었다 [1].

이와 같은 분석은 단순한 성능 비교가 아니라, 워크로드 - 가속기 매핑 전략을 제시한다는 점에서 중요하다. 즉, CNN 기반 태스크는 FPGA/ASIC에, BEV 기반 인식은 GPU/NPU에, trajectory prediction은 NPU/ASIC에 최적화된다. 이를 정리한 것이 표 1과 같이, 자율주행 주요 태스크별로 적합한 하드웨어를 대응시킨 매핑 전략이다. 이러한 전략은 향후 chiplet SoC 설계나 in-memory computing 연구에서 실제 하드웨어 아키텍처 선택에 지침을 제공할 수 있다 [1,6].

V. 결론

트랜스포머 모델은 자율주행 인식과 예측 성능을 크게 향상시키며, BEVFormer나 TransFusion과 같은 최신 접근법은 복잡한 주행 환경에서도 강력한 성능을 보여준다 [5,6]. 그러나 이러한 워크로드는 높은 연산량과 메모리 대역폭을 요구하여, 차량 내 임베디드 플랫폼에서 실시간 실행 시 전력·발열·지연의 병목이 발생한다 [1]. 본 논문은 자율주행 파이프라인 속 트랜스포머 응용을 정리하고, GPU, FPGA, ASIC/NPU와 같은 가속기의 역할 및 태스크별 매핑 전략을 검토하였다. 향후 연구는 chiplet 기반 SoC, in-/near-memory computing, 분산 추론 전략을 통해 이러한 제약을 완화하고, 실세계 자율주행 시스템에서 트랜스포머 추론의 효율성과 상용화 가능성을 더욱 높일 것으로 기대된다 [3,6]

Autonomous Driving Task	Workload Type	Best-fit Accelerator	Rationale
Lane/Traffic Light Detection	CNN (regular conv)	FPGA / ASIC	Structured pipelines, low power, deterministic latency [5]
BEV Construction (BEVFormer, BEVFusion)	Transformer (attention-heavy)	GPU / NPU	Parallel tensor ops, high throughput [6]
Trajectory Prediction	Transformer (seq2seq)	NPU / ASIC	Deterministic low-latency inference [6]
Multi-sensor Fusion(LiDAR+Camera)	Hybrid CNN+Transformer	GPU + NPU	Balanced workload split, memory bandwidth critical [3]

표 1. 자율주행 워크로드-가속기 맵핑

참 고 문 헌

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, “A survey of autonomous driving: Common practices and emerging technologies,” *IEEE Access*, vol. 8, pp. 58443 - 58469, May 2019.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 3213 - 3223.

[3] Y. Zhou and O. Tuzel, “VoxelNet: End-to-end learning for point cloud based 3D object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, July 2017, pp. 4490 - 4499.

[4] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, and Z. Chen, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, June 2020, pp. 2443 - 2451.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137 - 1149, June 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, December 2017.

ACKNOWLEDGMENT

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2024-00436887). 본 논문의 교신 저자는 김중헌임.