카메라-레이더 센서 퓨전 성능 향상을 위한 Tabular Transformer 기반 레이더 데이터 통합 방안

박재현, 김성철
서울대학교 전기정보공학부 뉴미디어 통신 공동 연구소
{ltbljb1677, sckim} @ maxwell.snu.ac.kr

Leveraging Tabular Transformers for Enhanced Radar Metadata Integration in Camera-Radar Sensor Fusion Tasks

Jae-Hyun Park, Seong-Cheol Kim
Department of Electrical and Computer Engineering and INMC, Seoul National. Univ.

요 약

자율주행 기술은 차량 운행의 효율성과 승객 안전을 보장하기 위해 주변 환경을 정확하게 인식하고 해석하는 능력에 크게 의존한다. 이러한 능력은 자율주행 차량에 장착된 센서들의 성능과 그 인지 기술에 기반하고 있다. 특히, 레이더와 카메라 센서의 융합은 광범위한 환경 조건 하에서 차량 인지 시스템의 견고성과 정확성을 향상시킬 수 있는 중요한 잠재력을 가지고 있다. 최근에는 3D 인식 기술이 전면 시야에서 Bird-Eye-View (BEV)로의 전환이 이루어지면서 물체 탐지 및 다양한 하위 작업의 성능이 크게 향상되었다. 본 논문에서는 SimpleBEV 모델을 변형하여 다중 모달 작업에 맞게 조정된 레이더 메타 데이터의 통합을 위해 Tabular Transformer를 적용한 새로운 접근법을 제시한다. 본 연구에서의 퓨전 모델은 레이더 메타 데이터 속성의 혼합 임베딩을 학습할 수 있도록 설계되었으며, Tab-Transformer와 FT-Transformer 모델을 결합하여 Rasterize Module을 구성하였다. 결과적으로 Baseline에 비해 IOU 성능이 2.9% point 개선되는 것을 확인할 수 있었다.

I. 서론

자율주행 분야에서 주변 환경을 해석하고 탐지하는 능력은 자율주행 차량의 효율성과 탑승객의 안전을 위해 중요하다. 이러한 능력은 자율주행차량에 부착된 센서들의 성능과 인지 기술에 기반하고 있으며 주로 객체 인식 (object detection), 객체 추적 (object tracking), 이미지 분할 (image segmentation) 등의 Task에서 측정된 성능을 토대로 평가된다. 사용되는 다양한 센서 중에서, 레이더와 카메라 센서의 융합 (fusion)은 광범위한 환경 조건에서 차량의 인지 시스템의 견고성과 정확성을 향상시킬 수 있는 잠재력을 지니고 있다. 특히 레이더 센서는 나쁜 날씨 조건에서의 신뢰성과 거리 및 속도를 측정할 수 있는 능력이 우수하고, 카메라는 고해상도 시각 데이터를 보완할 수 있다.

최근, Camera 기반 3D Perception 분야는 전면 시야 (Front-View)에서 Bird-Eye-View (BEV)로 전환함으로써 자율주행 환경에서 물체 탐지 성능과 다양한 Downstream Task의 성능의 향상을 이루었다 [1], [2], [3], [4]. 이러한 Paradigm Shift에 따라, Camera-Radar Sensor Fusion 분야 역시 BEV 상의 객체 검출 연구들이 활발하게 진행되고 있다. Camera-Radar Sensor Fusion에서 높은 성능을 보이는 다수의 연구들이 Camera 기반 3D Object Detection의 최신 연구들에서 설계된 모델을 차용해 Camera Branch로 사용하고, Radar Branch와 Fusion Operation을 설계하는 방식으로 멀티 모달 Task에 맞게 모델을 설계하였다 [5], [6].

이 연구에서는 최신 3D Perception 모델 중의 하나인 SimpleBEV 모델 [2]을 수정하여 19개 레이더 메타 데이터를 효율적으로 활용하는 것을 목표로 한다. 이 데이터는 레이더 포인트당 위치, 속도, RCS, 동적 상태 등의 정보를 포함한다. SimpleBEV 모델 [2]에서는 nuScenes 데이터셋의 레이더 포인트별 메타 데이터를 Stack해 BEV Coordinate에 Rasterize 한 후 Convolution Operation을 하는 Compressor를 통해 데이터를 통합하는 방식을 사용하였다. 그러나 레이더 메타 데이터는 Tabular 데이터 형식으로 Smooth Solution에 편향되는 경향이 있는 CNN-like inductive bias가 적절하지 않다는 사실이 보고되어왔다 [7]. 또한 Neural Net은 Tabular data의 특성상 다수 포함되어 있는 uninformative 데이터에 영향을 많이 받는다고 알려져 왔다. 따라서 본 연구에서는 적절한 Inductive Bias를 활용하여 레이더 메타 데이터를 통합하기

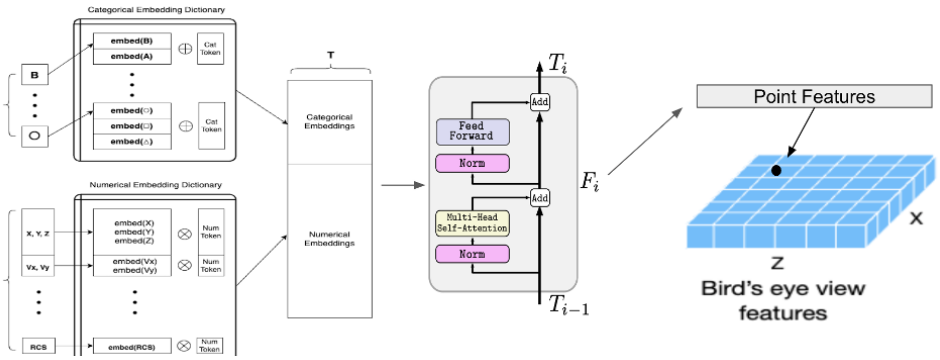

그림 1. Transformer 기반 Rasterize 모듈

위해 Tabular Data를 토대로한 최신 모델인 Tab-Transformer [8]와 FT-Transformer [9]모델을 수정하여 레이더 정보를 BEV Coordinate 상에 Mapping 하는 Rasterize Module을 설계하였다.

II. 본론

2.1 SimpleBEV

SimpleBEV [2]는 unprojection을 통해 6개 View의 Image Feature를 BEV Coordinate에 대응하는 Rasterizing 과정을 거친다. 이때 Bilinear Interpolation을 통해 대응되는 Feature를 BEV Coordinate 상에 Mapping 한다. Radar Meta Data를 포함한 Radar의 Point Cloud를 동일하게 BEV Coordinate 상 Mapping하고, Camera Branch에서 얻은 Image Feature와 Radar Branch의 Radar Feature를 그림 2와 같이 Concatenation 한다. 그 후 Convolution Operation을 토대로 Image와 Radar Feature에서 Latent Feature를 얻고, Latent Feature를 U-Net에 통과시켜 BEV 상에서 Segmentation Task를 수행한다.

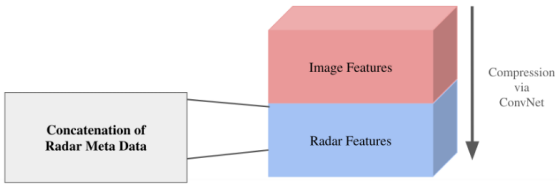

그림 2. SimpleBEV의 Fusion 과정

2.2 Rasterizing Module 설계

레이더 메타 데이터를 Integration 하기 위해 표1과 같은 레이더 메타 데이터를 Categorical 데이터와 Numerical 데이터로 나누었다. 가령 Id, Is_quality_valid, Ambig_state와 같이 사전에 범주가 정의된 데이터의 경우 Categorical 데이터로 분류하고, Coordinate value, Velocity 등과 같이 연속적인 값을 갖는 데이터는 Numerical 데이터로 분류한다.

Categorical 데이터의 경우 같은 Category 내에서 공유되는 Categorical Token과 Learnable Embeddings가 더해져 Encoding이 되고, Numerical 데이터는 Numerical Token이 Learnable Embeddings에 곱해져 Encoding이 수행된다. 결과적으로 얻어진 Embeddings가 Tabular Transformer의 Input으로 입력돼 레이더 메타 데이터의 Feature들 간의 Mixing이 일어난다. 이때, 레이더 메타 데이터 사이의 Channel Mixing과 Point Cloud 간의 Spatial Mixing이 동시에 수행된다. 결과적으로 얻어진 d-dimension의 Point Feature가 BEV 상에 Mapping 된다.

Attributes	Meaning
X, Y, Z	Coordinate value of radar point
dyn_prop	Moving attribute
Id	Point identifier
Rcs	Radar cross section
Vx, Vy	Velocity
…	…
Pdh0	False alarm rate
Vx_rms, Vy_rms	V Rms error
Time_delay	Time delay between measured and Reference Time

표 1 : 레이더 메타 데이터

2.3 실험 결과

Model	IOU (%)
Baseline : SimpleBEV	44.7
Baseline + Tabular Transformer (depth = 6, head num = 8)	45.9
Baseline + Tabular Transformer (depth = 4, head num = 4)	47.8

본 연구에서 설계한 Rasterize Module을 토대로 실험을 진행한 결과 Baseline SimpleBEV에 비해 2.9%의 IOU 성능 향상이 있음을 확인했다. 모델의 Depth와 Head 개수를 늘렸을 때 Rasterize Task에 비해 모델의 Capacity가 증가해 Overfitting이 발생하는 것을 확인할 수 있었다.

III. 결론

BEV 관점에서 레이더 메타 데이터를 Tabular Transformer를 사용하여 통합하는 접근 방식은 자율 주행 시스템의 인식 능력을 향상시키는데 유망한 결과를 보여주었다. SimpleBEV 모델에서는 레이더 메타 데이터를 통합하는 방식에 Convolution Operation을 사용했으나 Tabular Data인 특성상 Convolution이 적합하지 않음이 보고되어왔다. 이에 따라 본 연구에서는 Tabular Data에 적합한 Tabular Transformer를 토대로 레이더 메타 데이터를 인코딩해 Channel, Spatial Mixing을 진행했다. 결과적으로 Baseline 모델에 비해 IOU Score가 2.9% point 향상되었음을 확인하였다.

ACKNOWLEDGEMENT

이 연구는 (2024)년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20014098).

참 고 문 헌

[1] Philion, Jonah, and Sanja Fidler. "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer International Publishing, 2020.

[2] Harley, Adam W., et al. "Simple-bev: What really matters for multi-sensor bev perception?." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[3] Li, Yinhao, et al. "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 2. 2023.

[4] Li, Zhiqi, et al. "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022.

[5] Kim, Youngseok, et al. "Crn: Camera radar net for accurate, robust, efficient 3d perception." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[6] Lei, Kai, et al. "Hvdetfusion: A simple and robust camera-radar fusion framework." *arXiv preprint arXiv:2307.11323* (2023).

[7] Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." Advances in neural information processing systems 35 (2022): 507-520.

[8] Huang, Xin, et al. "Tabtransformer: Tabular data modeling using contextual embeddings." *arXiv preprint arXiv:2012.06678* (2020).

[9] Gorishniy, Yury, et al. "Revisiting deep learning models for tabular data." *Advances in Neural Information Processing Systems* 34 (2021): 18932-18943.