CLDS : 클래스별 Prototype 기반 객체 검출 성능 개선에 대한 연구

엄상우, 김지연, 전창재*
세종대학교

sangwoo4876@naver.com, lynnji2@naver.com, *cchun@sejong.ac.kr

CLDS : Improving Object Detection Performance via Class-wise Prototype

Sangwoo Eom, Jiyoun Kim, Chang-Jae Chun*
Sejong Univ.

요 약

최근 자율주행 기술의 발전과 함께 3 차원 객체 탐지 기술이 활발히 연구되고 있다. 이러한 멀티모달 객체 탐지는 서로 다른 센서 데이터를 결합하는 과정이 요구되며, 주로 각 센서로부터 추출된 정보를 공통 공간으로 정렬한 뒤 객체를 매칭하는 방식이 사용된다. 그러나 공통 공간 기반의 객체 매칭 방식은 표현 학습을 객체 단위에 집중시키는 경향이 있으며, 이로 인해 클래스 간 공통 특성을 학습하는데 구조적인 제약이 발생한다. 본 논문에서는 이러한 한계를 보완하기 위해 TransFusion 객체 탐지 모델에서 각 클래스의 대표적인 특징을 학습 과정에서 Prototype 형태로 형성하고, 이를 활용하여 객체 특징을 조절하는 방법을 제안한다. 실험 결과, 제안한 방법은 nuScenes 데이터셋 환경에서 기존 TransFusion 대비 전반적인 성능 저하 없이 Trailer 와 같은 일부 희소 클래스에 대해 탐지 평균 정밀도가 향상됨을 확인하였다.

Ⅰ. 서 론

최근 자율주행 기술의 발전과 함께 3 차원 객체 탐지 방법이 활발히 연구되고 있다. 이러한 멀티모달 객체 탐지는 서로 다른 형태의 센서 데이터를 결합하는 과정이 요구되며, 주로 각 센서로부터 추출된 정보를 공통 공간으로 정렬한 뒤 객체를 매칭하는 식이 널리 사용된다. 그러나 공통 공간 기반의 객체 매칭 방식은 클래스 간에 공유되는 공통 특성을 구조적으로 학습하는 데에 제약이 발생할 수 있다. 연구 [1]에서는 카메라와 LiDAR 로부터 추출된 특징을 BEV(Bird's Eyes View) 공간으로 정렬하여 융합하는 방법을 제안함으로써 센서 간 공간적 정합성을 개선하였다. 연구[2]에서는 Transformer 기반 구조를 활용하여 LiDAR 기반 객체 특징과 카메라 정보를 결합함으로써 객체 단위의 특징 상호작용을 효과적으로 모델링하였다. 그러나 두 방법 모두 개별 객체 중심의 표현 학습에 기반하고 있어 클래스 간 공통 특성을 추론 과정에서 직접적으로 활용하는 데에는 한계가 있다. 본 논문에서는 이러한 문제를 보완하기 위해 TransFusion 객체 탐지 모델에서 각 클래스의 대표적인 특징을 Prototype 형태로 형성하고, 이를 활용하여 객체 특징을 조절하는 방법을 제안한다. nuScenes 데이터셋 환경에서 실험한 결과, 기존 TransFusion 대비 전반적인 성능 저하 없이 Trailer 와 같은 희소 클래스에 대해 탐지 평균 정밀도가 향상됨을 확인하였다.

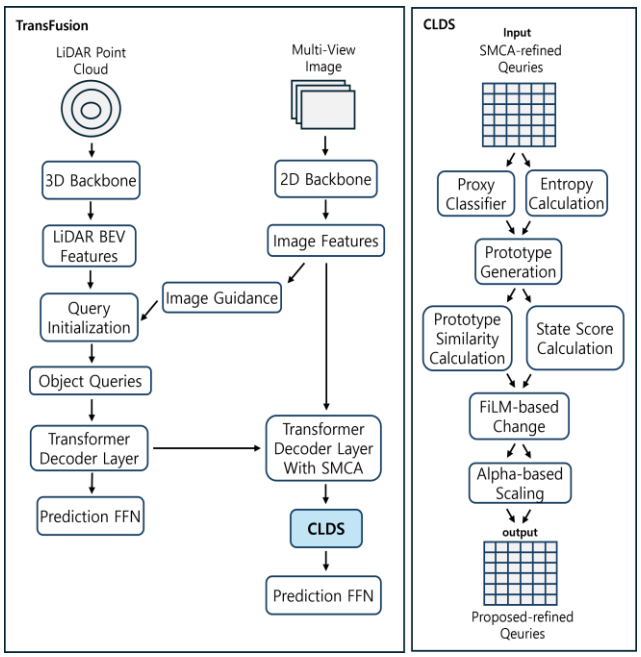
Ⅱ. 본론

2.1 데이터 구성

본 논문에서는 제안한 방법의 성능을 검증하기 위해 nuScenes[3] 데이터셋을 사용하였다. nuScenes 는 실제 도로 환경에서 수집된 대규모 자율주행 데이터셋으로, 카메라와 LiDAR 센서 정보를 포함한 멀티모달 데이터를 제공한다. 객체 탐지 실험에서는 nuScenes 에서 정의된 10 개 객체 클래스를 대상으로 학습 및 평가를 수행하였다.

2.2 제안 방법

TransFusion 은 카메라와 LiDAR 센서로부터 추출된 특징을 Transformer 구조를 통해 결합하는 3 차원 객체 탐지 모델이다. LiDAR BEV Feature 로 Object Query 를 생성한 뒤, Transformer Decoder 를 통해 LiDAR 기반 1 차 예측을 수행한다. 이후 예측된 3D 객체를 Image 공간에 투영하여 표현을 보정하며, LiDAR-Image Feature 를 결합하여 최종 객체 예측을 수행한다. 본 연구에서는 클래스별 특징을 Prototype 형태로 학습하여 객체 표현 보정에 사용하는 CLDS(Class-aware Latent Distribution Shaper)를 제안한다. 제안 방법은 LiDAR-only Decoder 출력과 Prediction Head 사이에 삽입되어, 객체 Query Feature 로부터 클래스 수준의 보조 표현을 도출하여 객체 표현을 조절한다. 이때 클래스 수준의 보조 표현은 각 클래스의 대표적인 특징을 요약한 Prototype 형태로 구성된다.



(a) (b)

그림 1. (a) 제안하는 모델 구조 및 (b) Class-aware Latent Distribution Shaper (CLDS) 구조

2.2.1 Class Prototype 생성

Prototype 개념은 각 클래스의 대표적인 특징을 요약된 형태로 표현하여 학습에 활용하는 방식으로, Prototypical Network[4]와 같은 Metric Learning 기반 분류 연구에서 효과적으로 사용되어 왔다. 제안 방법에서는 이러한 Prototype 개념을 재해석하여 각 클래스의 대표적인 특징을 Prototype 형태로 학습한다. 클래스별 Prototype 은 학습 가능한 임베딩 테이블로 구성되며, 디코더에서 출력된 객체 특징을 기반으로 해당 객체가 속할 가능성이 높은 클래스의 Prototype 을 참조한다. 객체 특징으로부터 Proxy Classifier 를 통해 클래스 확률을 추정하고, 클래스별 Prototype 을 가중합 형태로 혼합하는 Soft Prototype 방식을 적용한다. 이를 통해 객체의 불확실성을 고려한 부드러운 클래스 표현을 구성한다.

2.2.2 State 기반 Adaptive Gating

생성한 Prototype 을 그대로 반영할 경우 기존 Query Feature 를 과도하게 변화시켜 학습 안정성을 저해할 수 있다. 본 연구에서는 이러한 점을 고려하여 객체별 신뢰도를 반영한 State 기반 Adaptive Gating 을 도입한다. State 는 Prototype 과 객체 특징 간의 유사도, Heatmap 기반 분류 신뢰도, 그리고 클래스 예측의 엔트로피를 결합하여 객체별로 산출한다. 이 State 값은 Prototype 기반 변형을 객체 특징에 어느정도 반영할지를 결정하는 가중치로 사용된다. 이를 통해 Prototype 정보가 신뢰할 수 있는 경우에만 특징 조절이 강하게 적용되도록 한다.

2.2.3 Feature Modulation

Prototype 으로부터 유도된 특징 변형은 FiLM(Factorized Linear Modulation) 방식을 통해 객체 특징에 적용된다. FiLM 은 Prototype 을 입력으로 받아 객체 특징에 대한 Scale 및 Shift 파라미터를 생성하며, 원본 객체 특징에 잔차(Residual) 형태로 결합된다. 즉, 원본 특징과 Prototype 기반 변형 간의 차이를 State 기반 가중치와 함께 조절하여 기존 표현을 유지하면서도 클래스 정보를 점진적으로 반영한다.

2.3 실험 설정

본 논문에서는 제안한 Prototype 기반 Feature Modulation 모듈의 효과를 검증하기 위해 nuScenes 데이터셋을 사용하여 실험을 수행하였다. 모든 실험은 동일한 학습 및 평가 설정에서 진행되었으며, 기존 TransFusion 모델을 Baseline 으로 설정하여 성능을 비교하였다. 모델 성능 평가는 평균 정밀도(mean Average Precision, mAP)를 기준으로 수행하였다.

2.4 정량적 성능 비교

표 1 은 TransFusion 모델과 제안 방법인 CLDS 를 적용한 모델 간의 클래스별 mAP 를 비교한 결과이다.

모델 \ Class	Car	Truck	Bus	Trailer	Construction Vehicle	Pedestrian	Motorcycle	Bicycle	Traffic Cone	Barrier
TransFusion	87.8	63.9	74.2	43.2	29.3	88.1	74.4	64.4	77.2	69.5
Proposed	**87.8**	**64.1**	**74.3**	**44.2**	**29.9**	**88.2**	**74.6**	**64.6**	**77.3**	**69.5**

표 1. Class 별 AP

실험 결과, 대부분의 클래스에 대해서는 기존 모델 대비 mAP 가 유지되는 것을 확인하였으며, 희소 클래스인 Trailer 와 Construction Vehicle 에 대해서는 각각 1.0%p, 0.6 %p 의 mAP 향상이 나타났다. 이를 통해 모델이 클래스별 대표 특성을 Prototype 형태로 함께 학습하고 객체 표현 보정에 활용함으로써, 탐지 평균 정밀도가 향상되었음을 확인하였다.

Ⅲ. 결론

본 논문에서는 TransFusion 기반 3 차원 객체 탐지 모델의 객체 중심 표현 학습 특성을 분석하고, 클래스 수준 정보를 보다 효과적으로 활용하기 위한 새로운 특징 조절 방식을 제안하였다. nuScenes 데이터셋을 이용한 실험 결과, 대부분의 객체 클래스에 대한 mAP 는 유지하면서 Trailer 및 Construction Vehicle 과 같은 일부 희소 클래스에 대한 mAP 향상이 나타났다. 이를 통해 클래스 수준 정보를 활용한 표현 조절 방식이 데이터 분포가 제한적인 객체에 대해 보다 안정적인 탐지 성능을 유도할 수 있음을 확인하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었습니다(IITP-2026-RS-2023- 00254529).

참 고 문 헌

[1] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao wang, Tao Tang, Bing Wang, and Zhi Tang, "BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework" in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai, TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers" in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[3] Caesar, Holger, et al, "nuScenes: A multimodal dataset fot autonomous driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[4] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in Neural Information Processing System. 2017.