정확하고 강건한 3D 객체 검출을 위한 이중 SpConv 구조를 적용한 DSVT 기반 아키텍처

배상욱, 한동석*

경북대학교 전자전기공학부

sangukbae99@gmail.com, *dshan@knu.ac.kr

A DVST-Based Architecture with Dual SpConv Enhancement for Accurate and Robust 3D Object Detection

Sang Uk Bae, Dong Seog Han*

School of Electronic and Electrical Engineering, Kyungpook National Univ.

요 약

본 논문에서는 3차원 객체 탐지를 위한 DSVT(Dynamic Sparse Voxel Transformer) 기반 아키텍처를 제안한다. 기존 DSVT는 희소한 포인트 클라우드를 효율적으로 처리할 수 있는 어텐션(Attention) 기반 구조를 갖추고 있으나, 합성곱 연산을 통한 명시적 특징 정제가 부족하다는 한계가 있다. 이를 보완하기 위해 프리 어텐션 SpConv을 통해 지역적 기하 정보를 강화하고, 포스트 어텐션 SpConv을 활용하여 보다 넓은 문맥 정보를 반영하였다. 또한, 소규모 배치 환경에서도 안정적인 학습이 가능하도록 배치 정규화 대신 그룹 정규화를 적용하였다. 제안한 아키텍처는 기존 DSVT의 연산 효율을 유지하면서도 표현력, 탐지 정확도, 학습 안정성 측면에서 향상된 성능을 보인다. 실험은 nuScenes 데이터셋을 기반으로 수행되었으며, 원본 DSVT에 비해 정확도에서 우수한 결과를 나타낸다.

Ⅰ. 서 론

3차원 객체 탐지는 자율 주행, 로봇 비전, 드론 기반 감시 등 다양한 분야에서 핵심 기술로 활용되고 있다. 특히 라이다 센서로부터 획득한 포인트 클라우드는 장거리에서의 거리 정보를 정밀하게 제공하지만, 그 구조가 불규칙하고 희소하다는 특성으로 인해 고전적인 2차원 영상 처리 방식으로는 효과적인 특징 추출이 어렵다.

이러한 포인트 클라우드를 처리하기 위해 일반적으로 사용하는 방식이 복셀화이다. 이는 연속적인 3차원 공간을 격자 형태의 작은 정육면체 단위인 복셀로 나누고, 각 복셀 단위로 점 정보를 집계하여 처리함으로써 구조화된 입력 형태로 변환하는 과정이다. 복셀 기반 방식은 포인트 클라우드의 공간 정보를 보존하면서도 연산 효율을 확보할 수 있어 3차원 탐지에서 널리 활용된다[1].

DSVT(Dynamic Sparse Voxel Transformer)는 복셀 기반의 3차원 탐지 모델 중 하나로, 트랜스포머(transformer) 기반 어텐션 구조를 희소한 3차원 공간에 효과적으로 적용한 대표적인 모델이다[2]. 이 모델은 윈도우 단위로 복셀을 처리하는 구조를 통해 연산 효율성과 표현력을 동시에 확보하며, 포인트 클라우드의 불균형한 분포에 유연하게 대응할 수 있는 구조적 장점을 가진다. 또한, 연속된 어텐션 연산 간 정보 단절을 최소화하고, 인접 윈도우 간 문맥 정보를 효율적으로 전달할 수 있는 설계를 포함하고 있어 실시간 3차원 인식 환경에서도 우수한 성능을 보인다[2].

그러나 DSVT는 어텐션 기반 연산에 초점을 맞춘 나머지, 다음과 같은 한계가 존재한다[2]. 첫째, 명시적인 합성곱 연산이 부족하여 세밀한 기하 구조 표현이 제한되며, 입력 포인트 밀도가 낮거나 복잡한 구조일 경우 표현력이 약해질 수 있다. 둘째, 지역적 특징 보강 구조가 부재하여, 특히 작은 객체나 얇은 구조를 가진 객체에서 탐지 성능이 저하될 수 있다. 셋째, 배치 정규화에 의존하고 있어, 실사용 환경에서 자주 발생하는 소규모 배치 학습에 불안정성을 초래할 수 있다.

이러한 문제를 해결하기 위해 본 논문에서는 DSVT 기반 구조에 프리 어텐션 SpConv과 포스트 어텐션 SpConv을 각각 추가한 이중 정제 구조를 제안한다. 프리 어텐션 SpConv는 어텐션 이전에 지역적 기하 정보를 강화하며, 포스트 어텐션 SpConv는 팽창 합성곱(dilated convolution)을 통해 전역 문맥 정보를 보완한다[3]. 또한, 학습 안정성을 높이기 위해 배치 정규화 대신 그룹 정규화를 적용하였다.

제안한 구조는 nuScenes 데이터셋을 기반으로 평가하였으며, 기존 DSVT 대비 향상된 탐지 성능을 달성하였다. 어텐션 중심 구조와 SpConv 기반 정제 연산의 결합은 상호 보완적인 역할을 하며, 복잡하고 다양한 3차원 환경에서도 강건한 탐지가 가능함을 입증한다.

Ⅱ. 제안 방식

본 연구에서 제안하는 구조는 기존 DSVT의 희소 어텐션 기반 설계를 유지하면서, 그 한계를 보완하기 위한 두 가지 핵심 개선 요소를 중심으로 구성된다. 기존 DSVT는 윈도우 기반 어텐션 구조를 통해 희소한 포인트 클라우드에 내재된 지역 및 전역 정보를 효과적으로 통합할 수 있으며, 복잡한 커스텀 연산 없이도 TensorRT와의 호환성을 유지하는 실용적인 장점을 가진다[2]. 그러나 어텐션 중심 구조만으로는 세밀한 기하 구조를 안정적으로 포착하기 어렵고, 초기 복셀 특징의 표현력 부족이나 학습 안정성의 문제가 발생할 수 있다.

이러한 한계를 해결하기 위해, 어텐션 연산의 전후에 각각 프리 어텐션 SpConv과 포스트 어텐션 SpConv 구조를 도입하였다. 먼저 프리 어텐션 SpConv은 입력 복셀 특징에 대해 초기 단계에서 지역적 기하 정보를 강화하는 역할을 수행한다. 어텐션은 전역적인 관계 표현에는 유리하지만, 지역 필터링이나 작은 객체의 경계와 같은 세부 정보를 포착하는 데는 한

계가 있다. 이를 보완하기 위해 프리 어텐션 SpConv은 두 단계의 희소 합성곱 연산과 하나의 잔차 블록으로 구성된다[3]. 첫 번째 희소 합성곱은 입력 특징의 기초적인 공간 상호작용을 유도하고, 두 번째 합성곱은 이를 더욱 심화시키는 방식으로 설계되었다. 각 합성곱 연산 후에는 그룹 정규화와 비선형 활성화 함수(ReLU)를 적용하여 표현력을 향상시킨다. 이어지는 잔차 블록은 동일한 차원의 입력과 출력을 연결하여 네트워크의 수렴 속도와 안정성을 높이며, 어텐션 블록에 전달될 복셀 특징의 품질을 사전에 보강하는 역할을 수행한다.

어텐션 블록은 본 연구에서 구조 변경 없이 기존 DSVT의 설계를 그대로 따르며, 프리 어텐션 SpConv 및 포스트 어텐션 SpConv 블록과 결합되어 보완적 역할을 수행한다. 이 블록은 희소한 복셀을 윈도우 단위로 나눈 뒤, 각 윈도우 내에서 복셀 수를 기준으로 균일한 크기의 하위 집합으로 분할하여 어텐션 연산을 수행한다. 이를 동적 세트 분할이라 하며, 윈도우마다 포인트 밀도가 상이한 상황에서도 연산 단위를 균일하게 유지하고, 전체 어텐션 처리를 병렬화하는 데 기여한다[2].

또한, 어텐션 블록은 두 층으로 구성되며, 첫 번째 층에서는 X축 기준, 두 번째 층에서는 Y축 기준으로 복셀을 정렬하여 회전 기반 분할을 수행한다. 이 구조는 하위 집합 간의 연결을 강화하고, 단일 분할 방향에서 발생할 수 있는 정보 단절을 방지한다. 여기에 하이브리드 윈도우 전략이 더해져, 연속된 블록 간 윈도우의 크기나 위치를 교대로 변화시키면서 연산 효율을 유지하고, 더 다양한 공간 정보를 통합할 수 있도록 한다[2]. 이러한 설계를 통해 어텐션 블록은 희소한 포인트 클라우드에서도 지역적 세부 정보와 전역 문맥 정보를 효과적으로 학습할 수 있도록 지원한다.

어텐션 이후에는 포스트 어텐션 SpConv을 삽입하여 전체 문맥 정보를 보강한다. 이 구조는 팽창 합성곱을 적용한 희소 합성곱 연산과 하나의 잔차 블록으로 구성된다. 팽창 합성곱은 수용 영역을 확장하여 보다 넓은 범위의 공간 정보를 통합할 수 있게 하며, 이를 통해 어텐션으로 학습된 전역적 연결성 이후에도 지역적 패턴이나 세부 구조를 효과적으로 보완할 수 있다[3]. 이어지는 잔차 블록은 전처리 단계와 마찬가지로 특징 정제를 강화하고, 모델의 일반화 성능을 높이는 역할을 수행한다.

제안한 구조는 입력과 출력의 포맷을 기존 DSVT와 동일하게 유지하면서, 내부적으로는 특징 보강과 학습 안정성을 고려한 변화만을 적용하여 기존 구조와의 호환성 및 확장성을 확보하였다[2]. 이러한 설계는 어텐션과 SpConv의 장점을 결합함으로써, 복잡하고 다양한 기하 구조를 가진 3차원 객체를 보다 안정적으로 인식하고, nuScenes와 같은 실제 환경의 데이터셋에서도 정확도와 실시간성 측면에서 균형 잡힌 성능을 달성할 수 있도록 한다.

Ⅲ. 실험 및 평가

제안한 모델의 성능 평가는 nuScenes 데이터셋을 기반으로 수행하였다. 해당 데이터셋은 10개 클래스에 대한 3차원 라이다 기반 객체 정보를 포함한다. 실험에서는 백본(backbone) 구조에만 변화를 주고, 나머지 구성 요소는 기존 TransFusion과 동일하게 유지하였다. 3차원 트랜스포머 기반 백본은 총 4개의 어텐션 블록으로 구성되며, 각 블록은 128차원의 특징 채널을 사용한다. 각 어텐션 블록 사이에는 풀링 연산이 존재하지 않는 단일 스트라이드 구조이며, 포스트 어텐션 단계에서도 128차원을 유지하여 출력 BEV 맵의 채널 수 역시 128로 고정된다. 이후 2차원 백본은 128 - 128 - 256 구조의 피처 맵을 거쳐 검출기로 전달된다.

입력 포인트 클라우드는 [−54, −54, −5], [54, 54, 3] 범위를 커버하고, 복셀 크기는 [0.3, 0.3, 8.0]로 설정하여 360*360*1 크기의 희소 텐서 형태로 변환된다. 학습은 GPU당 배치 크기 2, 총 5에폭으로 진행되었으며, 최적화에는 Adam OneCycle 스케줄러와 초기 학습률 0.005를 사용하였다.

표 1. nuScenes 데이터셋을 이용한 결과표

Methods	mAP(↑)	NDS(↑)	mATE(↓)	mAOE(↓)
DSVT	0.4199	0.5103	0.3549	0.5505
Ours	**0.5227**	**0.5969**	**0.3247**	**0.4645**

표 1은 원본 DSVT 모델과 제안한 모델의 성능을 비교한 결과를 나타낸다. 제안한 모델은 모든 주요 지표에서 원본보다 상당히 우수한 성능을 보였다. 특히 mAP와 NDS에서 각각 약 +0.10 이상 상승하여 전반적인 탐지 정확도와 인식 신뢰도가 개선되었음을 보여준다. 또한, mATE와 mAOE 지표에서도 모두 오차가 줄어들어, 더 정확한 위치 예측과 안정적인 방향 추정이 가능함을 확인할 수 있다. 이러한 성능 향상은 어텐션 블록 전후에 삽입한 프리 어텐션 SpConv과 포스트 어텐션 SpConv이 지역적 기하 정보를 보강하고, 전역 문맥 정보를 보완한 결과로 해석된다. 나아가, 그룹 정규화 기법을 도입하여 소규모 배치 환경에서도 안정적인 학습이 가능하도록 설계된 점이 이러한 결과에 기여한 것으로 판단된다.

결과적으로, 제안한 백본 구조는 기존 DSVT의 효율성과 호환성을 유지하면서도, 실질적인 성능 개선을 효과적으로 달성하였다.

Ⅳ. Conclusion

본 연구에서는 DSVT의 희소 어텐션 구조에 프리 어텐션 SpConv 및 포스트 어텐션 SpConv 블록을 추가하여, 지역적 기하 정보와 전역 문맥 정보를 균형 있게 학습할 수 있는 3차원 인식 모델을 제안하였다. 제안한 구조는 기존 DSVT의 효율성과 배포 친화성은 유지하면서도, 복셀 간의 세밀한 상호작용과 특징 정제 과정을 보강함으로써 보다 정밀한 3차원 객체 탐지가 가능하도록 하였다. 본 연구는 트랜스포머 기반 백본 구조의 표현력을 강화하는 새로운 방향을 제시하며, 향후 다양한 3차원 인식 모델의 설계에도 유용한 기준점을 제공할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 국방기술진흥연구소에서 지원하는 국방 중소기업 역량강화 사업(No.DC2023CS)의 연구수행으로 인한 결과물임. This study is the result of the reserch performance of "Defense SMEs Competency Enhancement Program" (NO.DC2023CS) project supported by "Korea Research Institute for defense Technology planning and advancement"

참 고 문 헌

[1] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. "PointPillars: Fast Encoders for Object Detection from Point Clouds," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12697 - 12705, 2019.

[2] Xu, H., Zeng, Y., Shi, S., & Wang, Z. "Dynamic Sparse Voxel Transformer with Rotated Sets for 3D Object Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15574 - 15583, 2023.

[3] Chen, Y., Li, B., Zhang, Y., & Wang, Z. "VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking," arXiv preprint arXiv:2303.11301, 2023. (http://arxiv.org/abs/2303.11301)