3D 포인트 클라우드와 텍스트 연관성에 관한 연구

김다영[1], 조영준[2]
[1]전남대학교 소프트웨어공학과 학부생
[2]전남대학교 인공지능융합학과 교수
01055190398a@gmail.com, yj.cho@jnu.ac.kr

A Study on the Correlation between 3D Point Clouds and Text

Da-Yeong Kim[1], Yeong-Jun Cho[2]
[1]Dept. of Software Engineering, Chonnam National University
[2]Dept. of AI Convergence, Chonnam National University

요 약

최근 몇 년의 3D, 텍스트의 멀티 모달 연구의 방향성을 파악하고, 3D 포인트 클라우드 데이터와 텍스트 사이의 연관성을 파악할 수 있는 새로운 방법론을 제시한다.

1. 서론

최근 몇 년의 연구를 보면, 컴퓨터 비전 분야와 자연어 처리 분야를 통합하고자 하는 멀티 모달 모델이 많이 연구되어 왔고 급격한 발전을 이루어냈다. 해당 분야를 가속화 시킨 논문은 CLIP[1]으로, 텍스트 인코더, 이미지 인코더를 대용량 데이터셋으로 학습하여, Text와 Image 간의 연관성을 점수로 나타낸다. 하지만 아직 3D와 Text와의 연관성을 학습하는 것에는 아직 한계가 있다. 본 논문은, 2D 이미지 렌더링이나 복셀화에 의존하는 기존의 연구방식과 달리, 3D Patch를 이용해서 3D point를 직접 처리한다.

2. 소개

3D와 자연어의 통합은 멀티 모달의 핵심 과제로 부상하고 있다. CLIP [1]과 DALL-E [2] 같은 시각-언어 모델이 주목할 만한 발전을 이룬 반면, 이러한 아이디어를 3D 영역으로 확장하는 것은 여전히 뒤처져 있다. 이러한 격차는 주로 3D 데이터 표현의 본질적인 복잡성과 대규모 3D-텍스트 쌍 데이터셋의 부족 때문이다. 3D-텍스트 학습에 대한 기존 접근 방식은 종종 중간 2D 표현 또는 복셀화에 의존하는데, 이는 세밀한 3D 정보의 손실을 초래할 수 있다. 포인트 클라우드를 직접 처리하는 다른 방법들 은 장거리 의존성을 효율적으로 포착하는 데 어려움을 겪는다.

본 논문에서는 이러한 한계를 해결하기 위해 새로운 프레임워크를 제안한다. 프레임워크는 2D 이미지 처리에서 Vision Transformer [3]의 성공에서 영감을 받아 포인트 클라우드에 대한 패치 기반 인코딩 방식을 도입한다.

1. 정보 손실 없이 원시 3D 데이터를 효율적으로 처리하는 패치 기반 point cloud 인코더
2. 3D-텍스트 표현 학습에 맞춤화된 수정된 트랜스포머 아키텍처
3. 3D와 텍스트 표현을 공동 임베딩 공간에서 정렬하는 대조 학습 접근법

3. 관련 연구 분석

3.1 Image-Text Model

CLIP [1]의 성공적인 이미지-텍스트 결합 표현 학습은 3D 영역에서 여러 연구를 촉발시켰다. 3D-CLIP [4]은 CLIP을 3D 형상으로 확장했지만 다중 뷰 2D 렌더링에 의존한다. PointCLIP [5]은 CLIP의 이미지 인코더를 포인트 클라우드에 적용했지만 3D 구조 정보를 완전히 활용하지는 못한다. 본 논문은 사전 훈련된 2D 모델에 의존하지 않고 3D point cloud에서 직접 표현을 학습한다는 점에서 차이가 있다.

3.2 Vision Transformer

트랜스포머는 3D 이해 작업에서 유망한 결과를

보여주고 있다. Point Transformer[6]는 포인트 클라우드에 자기 주의 메커니즘을 적용했지만 패치 기반 접근 방식의 효율성이 부족하다. PCT는 분류와 분할을 위한 포인트 클라우드 트랜스포머를 도입했지만 3D 표현과 텍스트를 정렬하는 과제를 다루지 않았다.

4. 방법론

본 논문의 프레임워크는 세 가지 주요 구성 요소로 이루어져 있다: (1) Patch-based Point Cloud Encoder, (2) Text Encoder, (3) Contrastive Learning

4.1 Patch-based Point Cloud Encoder

N개의 점으로 구성된 입력 포인트 클라우드 $P \in \mathbb{R}^{(N\times 3)}$가 주어졌을 때, 이를 K개의 점을 포함하는 M개의 겹치지 않는 패치로 분할한다. 이 분할은 FPS(Farthest Point Sampling)과 KNN(K Neighbor Nearest) 그룹화를 통해 이루어진다. 각 패치는 PointNet 유사 네트워크를 사용하여 인코딩된다:

$$E_{patch}(P_i) = MLP(\max\{MLP(p_j) | p_j \in P_i\})$$

여기서 Pi는 i번째 패치이고, MLP1과 MLP2는 다층 퍼셉트론을 나타낸다. 이 패치 인코딩은 지역 기하 정보를 보존하면서 시퀀스 길이를 줄인. 인코딩된 패치들과 학습 가능한 class token을 결합하여 트랜스포머 인코더에 입력한다:

$$Z = TransformerEncoder([CLS; E_{patch}(P_1); \ldots; E_{patch}(P_M)])$$

트랜스포머 인코더는 Multi-Head Self Attention과 FeedFoward 네트워크로 구성되어 있어, 모델이 패치 간의 복잡한 관계를 포착할 수 있게 한다.최종적으로, class token에 해당하는 출력 벡터를 포인트 클라우드의 전체 표현으로 사용한다. 최종적으로, class token에 해당하는 출력 벡터를 포인트 클라우드의 전체 표현으로 사용한다:

$$F_{pc} = Z[0]$$

3.2 Text Encoder

텍스트 설명을 인코딩하기 위해, 우리는 CLIP기반 텍스트 인코더를 사용한다. 입력 텍스트 T 가 주어졌을 때, 텍스트 인코더는 고정 차원의 임베딩을 생성한다:

$$E_{text}(T) = CLIPTextEncoder(T)$$

3.3 Contrastive Learning

3D와 텍스트 표현 정렬을 위해, 대조 학습을 사용한다. N개의 (포인트 클라우드, 텍스트) 쌍으로 구성된 배치가 주어졌을 때, 모든 3D와 텍스트 임베딩 쌍 사이의 코사인 유사도를 계산한다:

$$S_{ij} = \cos(F_{pc}, E_{text})$$

양방향 대조 손실을 사용함으로써, 포인트 클라우드에서 텍스트로의 매칭과 텍스트에서 포인트 클라우드로의 매칭을 모두 고려한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었으며(IITP-2023-RS-2023-00256629) 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업의 연구결과로 수행되었습니다. (No. 2021-0-01409)

참고문헌

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[2] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International conference on machine learning (pp. 8821-8831). Pmlr.

[3] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Hegde, D., Valanarasu, J. M. J., & Patel, V. (2023). Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2028-2038).

[5] Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., ... & Li, H. (2022). Pointclip: Point cloud understanding by clip. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8552-8562).

[6] Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. (2021). Point transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16259-16268).