SemanticPointCLIP: Vision-Language for Point Cloud Semantic Understanding

Description

The use of vision-language models (VLP) has attracted significant attention in various communities due to their numerous applications. The CLIP model [1] was recently proposed as a contrastive learning approach that connects image and text and can be further utilized in many other applications [2, 4]. CLIP is trained on a large dataset and can be used for zero-shot learning on a new dataset. The image and text encoders can be utilized in other tasks, such as object detection and segmentation. While these approaches are primarily used in 2D images, point clouds are also essential in many areas, including SLAM and robotics manipulation. The most practical application requires the geometry structure for better inferring and interaction, especially in some industrial applications. Most existing works lack semantic understanding in a complex point cloud scenario. Combining scene understanding and natural language understanding [3] can greatly improve the performance of robotics applications. This work aims to use CLIP as the backbone for point cloud applications and bridge the gap between image and point cloud data, and transfer knowledge from images to the point cloud. We plan to utilize the pre-trained CLIP model, with task special prompt to build a point cloud semantic network. The network will then be able to segment the corresponding object with its prompts.

Tasks

- Literature review of vision-language models.
- Collection of point-text pairs for training and testing.
- Design a vision-language model structure for point cloud applications.
- Evaluate the proposed model using the prepared dataset.
- Compare the proposed network with other state-of-the-art approaches.
- Optional: submit the result to a top conference

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

Technische Universität München

TUM School of Computation, Information and Technology

Lehrstuhl für Robotik, Künstliche Intelligenz und Echtzeitsysteme

Supervisor:
Prof. Dr.-Ing. Alois Knoll

Advisor:
Jianjie Lin, M.Sc.

Research project:
SemanticPointCLIP

Type:
BA/MA

Research area:
CLIP, PointCloud, Semantic, Detection

Programming language:
Python

Required skills:
Very good mathematical background, programming in Python

Language:
English

For more information please contact us:

E-Mail: jianjie.lin@tum.de

Internet: www6.in.tum.de