Reaction report for “CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP”

Prachi Jain

What I like about this paper: I like that the paper is very innovative and explores how CLIP knowledge can benefit 3D scene understanding. CLIP is a powerful model that can learn from natural language supervision and perform zero-shot and few-shot learning on 2D images. However, applying CLIP to 3D point clouds is not straightforward and requires novel techniques to bridge the gap between 2D and 3D modalities. The paper makes the first attempt to investigate this problem and proposes a framework that transfers CLIP knowledge to a 3D point cloud network. Also, I like that the paper proposes a simple yet effective framework that leverages semantic and spatial-temporal consistency regularization to pre-train a 3D network. The framework consists of two components: a semantic-driven cross-modal contrastive learning module and a spatial-temporal consistency regularization module. The former uses CLIP's text semantics to select positive and negative point samples and then employs a contrastive loss to train the 3D network. The latter forces the consistency between the temporally coherent point cloud features and their corresponding image features. These two modules work together to transfer CLIP knowledge from 2D images to 3D point clouds in a self-supervised manner. Finally, I like the fact that the paper demonstrates impressive results on various downstream tasks and datasets and shows the potential of CLIP for 3D applications. The paper evaluates the pre-trained 3D network on three benchmarks: SemanticKITTI, nuScenes, and ScanNet. The paper shows that the pre-trained network achieves annotation-free 3D semantic segmentation with 20.8% and 25.08% mIoU on nuScenes and ScanNet, respectively, which are the first results of this kind. The paper also shows that when fine-tuned with 1% or 100% labelled data, the pre-trained network significantly outperforms other self-supervised methods, with improvements of 8% and 1% mIoU, respectively. Furthermore, the paper demonstrates the generalizability of the pre-trained network for handling cross-domain datasets, such as transferring from nuScenes to SemanticKITTI.

What I don’t like about this paper: The paper relies on the availability of 2D images that are aligned with the 3D point clouds. This may not be always feasible or realistic in some scenarios, such as indoor scenes or occluded objects. It would be interesting to explore how to transfer CLIP knowledge to 3D point clouds without using 2D images or using fewer or noisy images. The paper uses a fixed set of text prompts to select the positive and negative point samples for contrastive learning. This may limit the diversity and richness of the semantic information that can be transferred from CLIP to the 3D network. It would be interesting to explore how to generate dynamic and adaptive text prompts that can better capture the semantic nuances and variations of the 3D scenes. The paper focuses on semantic segmentation as the downstream task, but there are other 3D scene understanding tasks that can benefit from CLIP knowledge, such as object detection, instance segmentation, panoptic segmentation, etc. It would be interesting to explore how to extend the framework to other tasks and evaluate the performance and generalizability of the pre-trained network.

Future directions: To transfer CLIP knowledge to 3D point clouds without using 2D images, one possible direction is to use a generative model that can synthesize realistic and diverse 2D images from 3D point clouds, and then use CLIP to align the generated images with the text semantics. Another possible direction is to use a cross-modal transformer that can directly encode and align the 3D point clouds and the text semantics without relying on 2D images. To generate dynamic and adaptive text prompts for contrastive learning, one possible direction is to use a natural language generation model that can produce relevant and diverse text prompts based on the 3D point cloud context. Another possible direction is to use a reinforcement learning model that can learn to select the optimal text prompts that can maximize the contrastive loss and the downstream task performance. To extend the framework to other 3D scene understanding tasks, one possible direction is to use a multi-task learning model that can jointly pre-train the 3D network on multiple tasks using CLIP knowledge. Another possible direction is to use a meta-learning model that can learn to adapt the pre-trained network to new tasks and datasets using few-shot or zero-shot learning techniques.